UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 14-12
Statistics and Econometrics Series 08
May 2014

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

IMPROVING THE GRAPHICAL LASSO ESTIMATION FOR THE PRECISION MATRIX THROUGH ROOTS OF THE SAMPLE COVARIANCE MATRIX

V. Avagyan[1], A. M. Alonso[2], and F. J. Nogales[3]

Abstract

In this paper, we focus on the estimation of a high-dimensional precision matrix. We propose a simple improvement of the graphical lasso framework (glasso) that is able to attain better statistical performance without sacrificing too much the computational cost. The proposed improvement is based on computing a root of the covariance matrix to reduce the spread of the associated eigenvalues, and maintains the original convergence rate. Through extensive numerical results, using both simulated and real datasets, we show the proposed modification outperforms the glasso procedure. Finally, our results show that the square-root improvement may be a reasonable choice in practice.

***Keywords*:** Gaussian Graphical Models, Gene expression, High-dimensionality, Inverse covariance matrix, Penalized estimation, Portfolio selection, Root of a matrix.

1 Department of Statistics, Universidad Carlos III de Madrid, Calle Madrid 126, 28903 Getafe (Madrid), Spain. Email: vaavagya@est-econ.uc3m.es.

2 Department of Statistics, Universidad Carlos III de Madrid, Calle Madrid 126, 28903 Getafe (Madrid), Spain. Email: andres.alonso@uc3m.es. Alonso gratefully acknowledge financial support from the Spanish Ministry of Science and Innovation grants ECO2011-25706 and ECO2012-38442.

3 Department of Statistics, Universidad Carlos III de Madrid, Avda. de la Universidad 30, 28911 Leganes (Madrid), Spain. Email: fjnm@est-econ.uc3m.es.

Improving the graphical lasso estimation for the precision matrix through roots of the sample covariance matrix

Vahe Avagyan* Andrés M. Alonso† Francisco J. Nogales ‡

May 10, 2014

Abstract

In this paper, we focus on the estimation of a high-dimensional precision matrix. We propose a simple improvement of the graphical lasso framework (glasso) that is able to attain better statistical performance without sacrificing too much the computational cost. The proposed improvement is based on computing a root of the covariance matrix to reduce the spread of the associated eigenvalues, and maintains the original convergence rate. Through extensive numerical results, using both simulated and real datasets, we show the proposed modification outperforms the glasso procedure. Finally, our results show that the square-root improvement may be a reasonable choice in practice.

Key Words: Gaussian Graphical Models, Gene expression, High-dimensionality, Inverse covariance matrix, Penalized estimation, Portfolio selection, Root of a matrix.

*Department of Statistics, Universidad Carlos III de Madrid, Calle Madrid 126, 28903 Getafe (Madrid), Spain. Email: vaavagya@est-econ.uc3m.es.

†Department of Statistics, Universidad Carlos III de Madrid, Calle Madrid 126, 28903 Getafe (Madrid), Spain. Email: andres.alonso@uc3m.es. Alonso gratefully acknowledge financial support from the Spanish Ministry of Science and Innovation grants ECO2011-25706 and ECO2012-38442.

‡Department of Statistics, Universidad Carlos III de Madrid, Avda. de la Universidad 30, 28911 Leganes (Madrid), Spain. Email: fjnm@est-econ.uc3m.es.

1 Introduction

In recent years, there has been a growth interest in estimating the inverse covariance matrix (also known as precision or concentration matrix) under a high dimensional setting. For instance, in finance an accurate precision matrix is required when computing optimal portfolios for a large number of assets (Frahm and Memmel (2010)). In machine or statistical learning methods, such as classification or clustering, a proper estimation of the precision matrix is fundamental when dealing with a vast amount of predictor variables (Mardia et al. (1979), McLachlan (2004)). Particularly important are those applications involving Gaussian Graphical Models (GGM), where the precision matrix is assumed to be sparse and its non zero entries are related with the partial correlation coefficients (Dempster (1972); Lauritzen (1996)). One notable application where the precision matrix is intrinsically sparse is the estimation of genetic regulatory networks through high dimensional microarray gene expression data (Stifanelli et al. (2013); Yin and Li (2013)). Other application involving sparse precision matrices is the estimation of functional brain connectivity networks through neuroimaging techniques (Huang et al. (2010)).

In this paper, we focus on the estimation of a high-dimensional precision matrix. There are several approaches that try to estimate efficiently such matrices. We assume a $n \times p$ *centered* sample data matrix, $\mathbf{X}$, is observed, where each row $X_i = (X_{i1}, ..., X_{ip})$ is a realization of a p-variate random vector that is independent and identically distributed for $i = 1, ..., n$, and has a covariance matrix Σ with the corresponding precision matrix $\Omega = \Sigma^{-1}$.

We first divide the approaches to estimate the precision matrix by considering those that estimate it by inverting an estimation of the covariance matrix, and those that estimate the precision matrix directly. We refer the former approaches as *two-step* estimation procedures where a covariance matrix must be estimated in the first step. The classical estimator of the covariance matrix is the sample covariance matrix S. However, when the ratio between the number of the variables p and the number of the observations n is less but close to one, then the bias of the corresponding inverse of the classical estimator may be large, $E(S^{-1}) = n/(n-p-2)\Omega$, and the associated precision matrix may be highly unstable. For instance, when $p = n/2 - 2$, then $E(S^{-1}) = 2\Omega$. Moreover, when $p/n > 1$, the classical estimator is not invertible. To overcome these difficulties, some approaches have been proposed to deal with the estimation of the covariance matrix when the dimension p is large compared with the number of observations n. All these approaches try to mitigate the effect of the smallest eigenvalues of the covariance matrix (see Chamberlain and Rothschild (1983), Bai (2003), Bai and Ng (2011)). One of the most successful approaches is the shrinkage estimator proposed by Ledoit and Wolf (2004) and extended by Schafer and Strimmer (2005). This estimator shrinks the sample covariance matrix toward a target matrix using a linear combination. But while this estimator presents good practical and theoretical properties, the associated inverse estimator may not inherit such properties. In particular, when the dimension of the problem is high, this inverse estimator may not be optimal and may amplify the estimation error of the covariance matrix estimator (Ledoit and Wolf (2012)). Moreover, these two-step approaches do not provide, in general, sparse precision matrix estimations. For these reasons, our proposed methodology will be based on the second class of approaches that attempt to estimate the precision matrix directly.

Following the ideas of Ledoit and Wolf (2004), a shrinkage approach can also be applied directly to the precision matrix estimation. In this way, Frahm and Memmel (2010) proposed a precision matrix estimation by considering a *convex* linear combination between the inverse of the sample covariance matrix and a target matrix. A similar study has been proposed by Kourtis et al. (2012), who considered a *conical* combination between the inverse of the sample covariance matrix and a target matrix. However, these two studies focus on reducing the out-of-sample variance of the portfolio returns, rather than obtaining a better precision matrix estimator. Furthermore, these two methods rely on the assumption that the ratio between the number of variables and the number of the observations is small enough ($p << n$).

Moreover, as explained previously, recent applications require the estimation of GGMs where conditional dependencies between the variables are estimated through the off-diagonal and nonzero entries of the precision matrix, which is assumed intrinsically sparse. To attain a sparsity pattern in the estimated precision matrix and deal with the case $p/n > 1$, the ℓ_1 or LASSO (Least Absolute Shrinkage and Selection Operator) regularization framework can be applied. This approach was proposed by Tibshirani (1996) in the regression framework. Banerjee et al. (2006) proposed the precision matrix estimation by maximizing the ℓ_1-penalized log-likelihood function to attain sparse solution. This approach has been extensively analysed by other authors[1] (e.g. Yuan and Lin (2007); d'Aspremont et al. (2008); Banerjee et al. (2008); Rothman et al. (2008); Yin and Li (2013)) and several efficient algorithms have been developed to solve problem efficiently, highlighting the Graphical LASSO (Friedman et al. (2008)), a Project Sub-gradient Method (Duchi et al. (2008)), an Alternating Linear Minimization (Scheinberg et al. (2010)), and an Interior Point method (Li and Toh (2010)), among others.

Regarding non-likelihood approaches, Meinhausen and Bühlmann (2006) proposed a neighborhood selection framework based on lasso regressions. Yuan (2010) proposed the use of the Dantzig selector to replace the lasso regression in this framework. Finally, Cai et al. (2011) introduced the constrained ℓ_1-minimization based on constraining the ℓ_1 norm of the precision matrix.

In this paper, we focus on the ℓ_1 penalized log-likelihood maximization approach and propose a simple modification that is able to attain a better statistical performance without sacrificing too much the computational cost. One of the most efficient algorithms to compute numerically the ℓ_1 penalized log-likelihood estimator is the *glasso* framework. This framework allows a fast, efficient and stable solution for high-dimensional problems. The glasso algorithm is based on minimizing the log-determinant of the precision matrix subject to its inverse is close to the sample covariance matrix, S. However, when p/n is large, it is well-known (Johnstone (2001)) that the eigenvalues of S are more spread and hence, its condition number is increased. To improve the stability of the glasso estimation, we propose to use a k-root of the sample covariance matrix, with $k \geq 1$, to attain less spread eigenvalues and therefore, obtain a more accurate estimation of $\Omega^{1/k}$ and also Ω.

The proposed k-root glasso algorithm is a simple modification of the glasso one: it is based again on minimizing the log-determinant of the precision matrix, but now subject to its k-root inverse is close to the k-root of the sample covariance matrix. Once the specific

[1]Other penalty functions have been proposed to regularize the log-likelihood, see Fan et al. (2009).

k-root and the penalty parameter (associated with the original glasso framework) are selected, the proposed procedure requires no additional cost than that of the glasso method. Through extensive numerical results, using both simulated and real datasets, we show the proposed technique outperforms the glasso estimator when considering different statistical losses and GGM performance measures. In particular, we use the entropy loss and the mean squared error to measure the statistical performance, and the specificity, the sensitivity and the Matthews Correlation Coefficient (MCC) to measure the GGM prediction accuracy. Furthermore, we propose an efficient calibration procedure for selecting the k-root of the sample covariance matrix and also the associated tuning (penalty) parameter that regularizes the log-likelihood. Finally, we show analytically that the convergence rate of the proposed k-root glasso remains the same as that of the glasso method.

The manuscript is organized as follows. Section 2 describes the proposed k-root glasso (or simply r-glasso) methodology to estimate large precision matrices. Section 3 analizes the convergence rate of the proposed framework. Section 4 proposes different efficient approaches for selecting both the k-root of the sample covariance matrix and the associated penalty parameter that regularizes the log-likelihood. Section 5 exhaustively evaluates the statistical loss and GGM performance of the proposed methodology and compare with that of the glasso one. Section 6 illustrates the solution properties when applying the proposed methodology to three empirical applications: the prediction of the breast cancer state, the prediction of the SRBC tumor, and the computation of an optimal financial portfolio. Section 7 gives the conclusions. Finally, Appendix A contains the proofs for the theoretical results in the paper and Appendix B illustrates the numerical results.

2 Proposed k-root glasso framework

Before proposing the k-root glasso methodology, we introduce the following notation. For any vector $\mathbf{a} = (a_1, ..., a_p)^T \in \mathbb{R}^p$, the ℓ_1 or Manhattan norm is denoted by $||\mathbf{a}||_1 = \sum_{j=1}^{p} |a_j|$, the ℓ_2 or Euclidean norm by $||\mathbf{a}||_2 = \sqrt{\sum_{j=1}^{p} a_j^2}$, and the ℓ_∞ or Maximum norm by $||\mathbf{a}||_\infty = \max(|a_1|, ..., |a_p|)$. For any symmetric matrix $\mathbf{A} = [a_{ij}]_{1 \leq i,j \leq p}$, the componentwise ℓ_1 norm is denoted by $||\mathbf{A}||_1 = \sum_{i=1}^{p} \sum_{j=1}^{p} |a_{ij}|$, the componentwise ℓ_2 or Frobenius norm by $||\mathbf{A}||_2 = \sqrt{\sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij}^2}$, the componentwise ℓ_∞ norm by $||\mathbf{A}||_\infty = \max_{1 \leq i,j \leq p} |a_{ij}|$, and finally the spectral norm by $||\mathbf{A}||_{\text{spec}} = \sup_{||x||_2 \leq 1} ||Ax||_2$. For any positive definite symmetric matrix A, $\lambda(A)$ will denote the vector of eigenvalues of matrix A, where $\lambda_{\max}(A) = \max \lambda_i(A) = ||A||_{spec}$ and $\lambda_{\min}(A) = \min \lambda_i(A)$ denote the maximum and minimum eigenvalues, respectively. Finally, we assume a centered sample data matrix, $\mathbf{X}$, is observed with dimension $n \times p$, where each row $X_i = (X_{i1}, ..., X_{ip})$ is a realization of a p-variate random vector that is independent and identically distributed as a multivariate normal for $i = 1, ..., n$, with covariance matrix Σ and precision matrix $\Omega = \Sigma^{-1}$.

The glasso estimator is defined as the solution of the following optimization problem:

$$\hat{\Omega}_{\text{glasso}} = \arg\max_{\Omega} \log\det(\Omega) - \text{trace}(S\Omega) - \nu||\Omega||_1, \tag{1}$$

where $S = (1/n)\sum_{i=1}^{n} X_i X_i^T$ is the sample covariance matrix and $\nu > 0$ is the penalty parameter. This parameter controls the sparsity pattern of the glasso estimation.

Note that problem (1) is convex, and its dual problem (2) is defined as (Banerjee et al. (2008)):

$$\hat{\Omega}_{\text{glasso}} = \arg\min_{\Omega} \ \log\det\Omega: \ \text{subject to } ||\Omega^{-1} - S||_\infty \leq \nu. \tag{2}$$

As commented in Section 1, the glasso estimation is sensitive to the eigenvalue structure of the sample covariance matrix, S, especially when the ratio p/n is large. To mitigate this sensitivity, we propose to shrink the eigenvalue spread by considering a k-root of S defined as $S^{1/k} = BV^{1/k}B'$, where $S = BVB'$ is the eigen-decomposition of S and $k > 1$. In this way, we propose the following k-root glasso estimator:

$$\hat{\Omega}_{\text{r-glasso}} = \arg\min_{\Omega} \ \log\det\Omega: \ \text{subject to } ||\Omega^{-1/k} - S^{1/k}||_\infty \leq \xi_k, \tag{3}$$

where $\xi_k > 0$ is the associated penalty parameter. Note when $k = 1$, the k-root glasso estimator reduces to the original one.

Again, problem (3) can be rewritten as

$$\hat{\Gamma} = \arg\min_{\Gamma} \ \log\det\Gamma: \ \text{subject to } ||\Gamma^{-1} - S^{1/k}||_\infty \leq \xi_k, \tag{4}$$

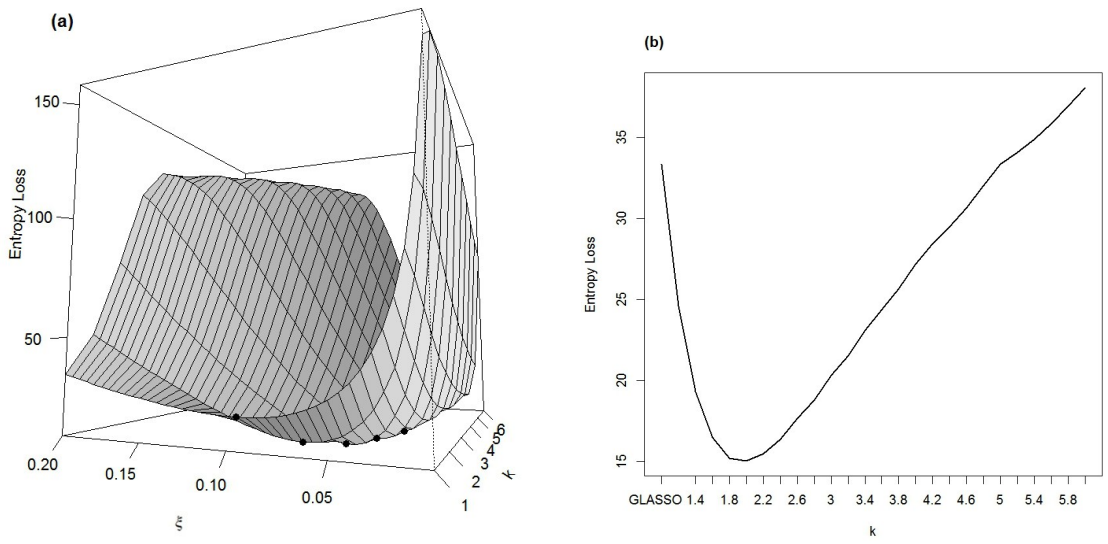
and we define the k-root glasso estimator as $\hat{\Omega}_{\text{r-glasso}} = \hat{\Gamma}^k$, for a given k and ξ_k. Note we can solve problem (4) using the same algorithm as for problem (2) without any additional cost.

To better understand the behavior of the proposed methodology, next we show a particular example. Assume the true precision matrix Ω and has the following sparse structure: $\omega_{ii} = 1$, $\omega_{i,i-1} = \omega_{i-1,i} = 0.5$, $\omega_{i,i-2} = \omega_{i-2,i} = 0.35$ and other elements are 0. For this example we choose the values $p = 300$ and $n = 500$.

In Figure 1(a), the entropy loss (see Section 5 for a formal definition) of the proposed estimator is shown as a function of different possible roots (between 1 and 6) and different values of the penalty parameter (between 0.005 and 0.2 with increments of 0.005). Note that, as the k-root moves away from 1 (which corresponds to the glasso estimator), it is possible to decrease the loss of the proposed estimator using convenient paths along ξ. That is, the minimum possible error of the glasso estimator along the ν path is larger than the minimum possible error of the proposed k-root glasso estimator along the ξ_k path. This improvement can be observed more clearly in Figure 1(b), where the entropy loss is plotted against k using the optimal value for ξ_k, that is, the penalty parameter that minimizes the entropy loss for a given k. Note we can reduce by half the statistical loss of the glasso estimator by using for instance the square-root glasso modification.

In Section 5, through an exhaustive empirical analysis including several sparsity patterns for the precision matrix, we confirm the proposed k-root glasso estimator may outperform the glasso one under other statistical performance measures covering those for graphical models.

Figure 1. **(a)** Entropy loss of $\hat{\Omega}_{\text{r-glasso}}$ as a function of ξ and k. **(b)** Entropy loss of $\hat{\Omega}_{\text{r-glasso}}$ as a function of k (given the optimal ξ_k).



3 Convergence Rate

In this section, we analyse the convergence rate of the proposed estimator $\hat{\Omega}_{\text{r-glasso}}$. Before proceeding with our results, the following main assumptions are made about the true precision matrix Ω:

A1 : $\lambda_{min}(\Omega) \geq \underline{\alpha} > 0$,

A2 : $\lambda_{max}(\Omega) \leq \bar{\alpha}$,

for some positive values $\bar{\alpha}$ and $\underline{\alpha}$.

Note that the assumptions A1 and A2 guarantee the existence of the matrix Ω. The following theorem presents the convergence rate of the proposed r-glasso estimator.

Theorem 1. *Suppose $\hat{\Omega}_{glasso}$ and $\hat{\Omega}_{r\text{-}glasso}$ are the solutions of problems (2) and (3), respectively. Under the assumptions A1, A2, the following result holds:*

$$||\hat{\Omega}_{r\text{-}glasso} - \Omega||_2 \overset{p}{\asymp} ||\hat{\Omega}_{glasso} - \Omega||_2, \tag{5}$$

where $A \overset{p}{\asymp} B$ means $A = O_P(B)$ and $B = O_P(A)$.

Proof. See Appendix A.

The Theorem 1 implies that the proposed estimator $\hat{\Omega}_{\text{r-glasso}}$ and glasso estimator $\hat{\Omega}_{\text{glasso}}$ have the same convergence rate under the Frobenius norm. This is an important results which shows that by improving the glasso estimator we do not impair the convergence rate of the estimator[2].

[2]For the convergence rate of the glasso estimator see Rothman et al. (2008)

4 Penalty Parameter Selection

The choice of the penalty parameter has a crucial role in all estimation procedures based on regularization. The penalty parameter controls the properties of the estimator, especially its sparsity level. To account for this sparsity level, we propose the use of the well-known BIC criterion,[3] proposed by Yuan and Lin (2007).

Our proposed methodology requires to calibrate two parameters ξ_k and k. We define the following BIC score to select simultaneously these parameters:

$$\text{BIC}(\xi_k, k) = n\left(-\log\det\hat{\Omega}(\xi_k, k) + \text{trace}(S\hat{\Omega}(\xi_k, k))\right) + \log(n) \times \text{NZ}, \tag{6}$$

where $\hat{\Omega}(\xi_k, k)$ is the estimated precision matrix using the values ξ_k and k, $\text{NZ} = \text{card}\{(i,j) : 1 \leq i \leq j \leq p, [\hat{\Omega}(\xi)]_{ij} \neq 0\}$ is the number of non zero elements of the matrix $\hat{\Omega}(\xi_k, k)$. The parameters (ξ_k, k) are selected by minimizing $\text{BIC}(\xi_k, k)$ using a grid search. Note that for $k = 1$ the value of (6) coincides with the value of the original BIC score function (see Yuan and Lin (2007)) .

5 Simulation Study

In this section, we perform a simulation analysis to compare the performance of the proposed estimator $\hat{\Omega}_{\text{r-glasso}}$ with that of the glasso one $\hat{\Omega}_{\text{glasso}}$. Specifically, in subsection 5.1 we detail the considered models for the precision matrix Ω, and in subsection 5.1 we describe the performance evaluation. Finally, in subsection 5.3 we provide the discussion of the results.

5.1 Considered models

We perform an exhaustive simulation study through eight different structures for the precision matrix with different sizes. We divide the models into random (where the sparsity pattern and the elements are chosen by chance) and non-random (with fixed sparsity pattern and deterministic elements). The considered models for the precision matrix Ω are the following:

(i) Random models

- *Model* 1. A random p.d. matrix, containing 5% of non-zero entries,
- *Model* 2. A random p.d. matrix, containing 10% of non-zero entries,
- *Model* 3. A random p.d. matrix, containing 20% of non-zero entries,
- *Model* 4. A random block-diagonal matrix, with four equally-sized blocks along the diagonal, each containing 50% of non zero entries.

(ii) Non-random models

[3] In one of the empirical applications in Section 6, we use a cross-validation procedure to calibrate the penalty parameter, as in this application the sparsity pattern is not important.

- *Model* 5. AR(1) structure: $\omega_{ii} = 1$, $\omega_{i,i-1} = \omega_{i-1,i} = 0.45$ and other values are 0 (Yuan and Lin (2007), Friedman et al. (2008)),
- *Model* 6. AR(2) structure: $\omega_{ii} = 1$, $\omega_{i,i-1} = \omega_{i-1,i} = 0.5$, $\omega_{i,i-2} = \omega_{i-2,i} = 0.35$ and other values are 0 (Yuan and Lin (2007)),
- *Model* 7. Decay structure: $\omega_{ij} = 0.6^{|i-j|}$ (Cai et al. (2011), Fan et al. (2009)),
- *Model* 8. A block-diagonal matrix, with four equally sized blocks along the diagonal, with a decay model in each block.

For each of the models, we simulate multivariate Normal random samples with zero mean and for $n = 500$ and $p = 100, 200, 300, 400, 500$. In each simulation, the number of replications is 100.

5.2 Performance evaluation

To compute the performance of a given estimator, $\hat{\Omega}$, we use the entropy loss function, also known as the Kullback-Leibler (KL) loss function, defined as follows:

$$\text{KLL}(\hat{\Omega}, \Omega) = \text{trace}(\Omega^{-1}\hat{\Omega}) - \log\det(\Omega^{-1}\hat{\Omega}) - p. \tag{7}$$

The KL loss function has been used widely in the literature involving covariance or precision matrices (see for instance Yuan and Lin (2007), Rothman et al. (2008), Fan et al. (2009), Yin and Li (2013)). Moreover, we also use the mean squared error defined as:

$$\text{MSE}(\hat{\Omega}, \Omega) = ||\hat{\Omega} - \Omega||_2^2. \tag{8}$$

Regarding the sparsity pattern or GGM prediction performance, we compute the specificity, sensitivity and Matthews Correlation Coefficient (MCC), defined as:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{9}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{10}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}, \tag{11}$$

where TP, TN, FP and FN are the numbers of true positives (number of correctly estimated non-zero entries), true negatives (number of correctly estimated zero entries), false positives (number of incorrectly estimated non-zero entries) and false negatives (number of the incorrectly estimated zero entries), respectively. Note that FP and FN can be seen as Type I and Type II errors, respectively. The MCC measure was introduced by Matthews (1975) and it is commonly used to measure the performance of binary classifiers. The MCC values are in [-1,1] and the closer the MCC to one, the better the classification.

Finally, we consider the glasso and the r-glasso procedures where the penalty parameters ν and ξ_k, and also the k-root parameter, are estimated using the BIC criterion (6). We also focus on the square-root glasso procedure, $k = 2$, because of its good behaviour in practice.

5.3 Discussion of results

The simulation results are provided in the Appendix B to conserve space. To better compare the performance of the proposed r-glasso estimator with that of the glasso one, we present the box plots of the loss differences, Δ(Loss). That is, for the KLL and MSE losses, we represent the difference between the corresponding estimation losses of glasso and r-glasso estimators. For the GGM performance measures, we represent the difference between the specificity, sensitivity and MCC of the r-glasso and glasso estimators. In all cases, positive values of the differences, Δ(Loss), will indicate the proposed r-glasso estimator outperforms the glasso one.

For precision matrices with random sparsity patterns (models 1-3), the differences Δ(Loss) are positive for KLL, MSE, specificity, sensitivity and MCC (see Figures 2-16). Therefore, the proposed estimation method outperforms glasso method in terms of statistical losses and prediction measurements. Similar results are obtained for block-type random precision matrix (model 4, see Figures 17-21).

For precision matrices with deterministic sparsity patterns, the results are similar for models 5 and 6. It can be seen the differences Δ(Loss) are positive for KLL (Figures 22, 27) and MSE (Figures 23, 28) and negative for specificity (Figures 24, 29) and MCC (Figures 26, 31). Therefore, the proposed estimation method outperforms glasso in terms of statistical losses, but glasso performs better than the proposed one in terms of GGM selection. In terms of the sensitivity, both methods perform identically (Figures 25, 30).

For model 7 and 8, the differences Δ(Loss) are positive for KLL, MSE, specificity, sensitivity and MCC[4] (see Figures 32-39). Therefore, the proposed estimation method outperforms glasso method in terms of the selected losses and prediction measurements.

Overall, the proposed r-glasso estimation method provides better performance than glasso estimation for most of the models. It dominates glasso method in terms of KLL and MSE for all the models. In terms of GGM selection, r-glasso dominates glasso especially for models with unknown or random sparsity patterns.

Finally, note that we obtain similar findings if we use the square-root glasso method ($k = 2$) instead of selecting the root ($k = k_{BIC}$) through the BIC criterion (6). This finding allows us to simplify and robustify our framework without sacrificing too much the performance.

6 Real Data Analysis

In this section we perform an empirical analysis of the proposed r-glasso method through three real-data applications: the first two based on predicting tumors while the last one based on selecting a large financial portfolio.

6.1 Breast Cancer Data

In this application, we focus on the problem of predicting breast cancer patients with pathological complete response (pCR). The literature has shown the pCR state after the neoadjuvant

[4]The specificity and MCC are excluded for model 7, because these measurements are not defined for dense models.

Table 1. Average pCR classification measurements over 100 replications.

Method	Specificity	Sensitivity	MCC
glasso	0.703	0.704	0.366
r-glasso $k = k_{BIC}$	0.686	0.828	0.450
r-glasso $k = 2$	0.695	0.820	0.452
r-glasso $k = 3$	0.690	0.826	0.452
r-glasso $k = 4$	0.683	0.828	0.446

chemotherapy strongly indicates a cancer-free life (Kuerer et al. (1999)). Thus, it is important to select the patients with the pCR state correctly. In our application we use a dataset containing gene expression levels[5], analysed previously by Hess et al. (2006). This dataset contains 22283 gene expression levels of 133 patients (subjects) with different stages of breast cancer. There are 34 patients with pCR and 99 patients with residual disease (RD).

First, we divide the data into a training set and a testing set with sizes 112 and 21, respectively. This process is repeated 100 times. We follow the same division scheme applied in Cai et al. (2011). The testing set randomly selects 5 subjects with pCR and 16 subjects with RD. The training set contains the remaining subjects. Second, for the training set we apply two sample t-test between the two groups, in order to select the most significant 113 genes with the smallest p-values. Finally, the precision matrix Ω is estimated with both the r-glasso and glasso methods, using the training set. The penalty parameters for both methods and the parameter k for r-glasso method are estimated using the BIC criterion, through two-dimensional grid search. Also we analyse the performance of the r-glasso method when the parameter k is selected from 2 to 4. The estimated precision matrix is used in the Linear Discriminant Analysis (LDA) score:

$$\delta_t(Y) = Y^T \hat{\Omega} \hat{\mu}_t - \frac{1}{2} \hat{\mu}_t^T \hat{\Omega} \hat{\mu}_t, \tag{12}$$

where $t = 1, 2$ (i.e. $t = 1$ for pCR and $t = 2$ for RD), $\hat{\mu}_t = \dfrac{1}{n_t} \sum_{i \in class_t} x_i$ is the within group average, calculated using the training data. The LDA score $\delta_t(Y)$ is used to classify the subject Y from the testing set. The rule for the classification is $\hat{t} = \arg\max \delta_t(Y)$ $(t = 1, 2)$.

To measure the prediction accuracy for the two methods, we use the specificity, sensitivity and the Matthews Correlation Coefficient (MCC), as defined in section 5.2. Moreover, we consider the TP and TN as the number of correctly predicted pCR and RD, respectively, and FP and FN as the number of falsely predicted pCR and RD, respectively. The average measures over 100 replications can be found in Table 1.

Note from Table 1, the proposed r-glasso using the BIC criterion has a higher MCC (more than 20%) than the glasso one, which indicates a better classification performance. Observe also that the proposed r-glasso method outperforms glasso based on sensitivity and obtain results based on specificity. Finally, note similar findings are obtained for r-glasso under other

[5]Available at http://bioinformatics.mdanderson.org/pubdata.html.

values of the root parameter k. Specifically, note that the square-root glasso ($k = 2$) is a good choice to improve performance.

As a robustness check, we have also repeated the same application by considering the most significant 200 genes instead of 113. The results can be found in Table 2.

Table 2. Average pCR classification measurements over 100 replications.

Method	Specificity	Sensitivity	MCC
glasso	0.750	0.628	0.350
r-glasso $k = k_{BIC}$	0.706	0.810	0.457
r-glasso $k = 2$	0.722	0.792	0.459
r-glasso $k = 3$	0.708	0.808	0.457
r-glasso $k = 4$	0.700	0.814	0.454

It can be seen the results are roughly similar. The classification measurement MCC shows that r-glasso provides a 30% improvement over glasso.

6.2 SRBC Tumor Data

In this application, we consider the problem of predicting the type of the Small Round Blue Cell (SRBC) tumors. The accurate prediction and diagnosis of the SRBC tumors is a major challenge, because the associated therapy and the treatment highly depend on the diagnosis (Khan et al. (2001)). We use the same dataset analysed by Khan et al. (2001), which contains expression levels of 2308 genes for 64 tissue samples[6]. In this dataset, there are four types of SRBC tumors: 12 tissues of Neuroblastoma (NB), 21 tissues of Rhabdomyosarcoma (RMS), 8 tissues of Burkit Lymphoma, a subset of non-Hodgkin Lymphoma (BL), and 23 tissues of Ewing family tumors (EWS).

First, we divide the data into a training set and a testing set with sizes 50 and 14, respectively. This process is repeated 100 times. To ensure that in both sets there will be tissues of all four types, we obtain the training set by randomly selecting 18 tissues from the EWS class, 6 tissues from BL class, 9 tissues from NB class and 17 tissues from RMS class (around 70% of the subjects from each class). The remaining 14 tissues will form the testing set. Second, we select the most significant 100 genes according to their F-statistics values. We rank the genes in the training set by the level of the information they provide using the F-statistics (Rothman et al. (2009)), defined as

$$F = \frac{\frac{1}{m-1}\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2}{\frac{1}{n-m}\sum_{i=1}^{m}(n_i - 1)s_i^2}, \tag{13}$$

where $m = 4$ is the number of the tumor classes, $n = 50$ is the number of the tissue samples, n_i is the number of the tissue samples of class i, $\bar{x}$ is the overall mean, $\bar{x}_i$ and s_i^2 are the sample

[6]Available at http://www.bioinf.ucd.ie/people/aedin/R/full_datasets/.

mean and the variance of the class i, respectively. Finally, using the training set, we estimate the precision matrix Ω both by r-glasso and glasso methods. The penalty parameters for both methods and the parameter k for r-glasso method are estimated using the BIC criterion, through two-dimensional grid search. Also we analyse the performance of the r-glasso method when the parameter k is selected from 2 to 4. The estimated precision matrix is used in the LDA score $\delta_t(Y)$, defined as (12), where $t = 1, 2, 3, 4$ is the index of tumor class. To measure the prediction accuracy, we use the average proportion of correctly classified tissues:

$$AP = \frac{1}{14} \sum_{i=1}^{100} NCC_i, \tag{14}$$

where NCC_i is the number of correctly classified tissues in the i-the replication. We also repeat the same application by considering the most significant 200 genes instead of 100. The results for both cases can be found in Table 3.

Table 3. The average proportion of correctly classified tissues over 100 replications.

Method	$p = 100$	$p = 200$
glasso	0.957	0.935
r-glasso $k = k_{BIC}$	0.990	0.984
r-glasso $k = 2$	0.989	0.983
r-glasso $k = 3$	0.990	0.985
r-glasso $k = 4$	0.990	0.985

Note from Table 3, the average prediction level is higher for the r-glasso estimator than that for the glasso one. The improvement rate of r-glasso over glasso is more than 3% when we consider the most significant 100 genes, and more than 5% when we consider the most significant 200 genes. Observe also that similar results are obtain if we select the square-root glasso ($k = 2$).

6.3 S&P 500 Portfolio Stock Selection

In our last application, we focus on developing a stock portfolio with minimum risk (variance). The precision matrix estimation plays a fundamental role in computing this optimal portfolio. It is well-known that the weights of the (global) minimum variance portfolio are defined as (see DeMiguel et al. (2009)):

$$w_{MVP} = \frac{\Omega \mathbf{1}_p}{\mathbf{1}_p' \Omega \mathbf{1}_p}, \tag{15}$$

where $\mathbf{1}_p$ denotes a vector of ones with length p. As the minimum-variance portfolio depends directly on the estimation of the precision matrix, an accurate estimation of such matrix may lead to decrease the out-of-sample risk or variance of the portfolio.

Following the empirical analysis by Goto and Xu (2013), we use monthly historical returns of the stock constituents of $S\&P$ 500 index for a total of $n = 240$ months.[7] We consider three

[7]The observations cover the period of April 1st 1994 - April 1st 2014.

different portfolios: a *small* portfolio with $p = 80$ of the largest stocks in the S&P 500 index, a *medium* portfolio with $p = 200$ randomly selected stocks and a *large* portfolio with $p = 300$ randomly selected stocks. To compute the estimated precision matrices, we apply the r-glasso and glasso methods using a "rolling-horizon" procedure as in DeMiguel et al. (2009) with 100 months, leaving 140 months to compute the out-of-sample portfolio variance of each procedure.

To estimate the penalty parameters for the precision estimation methods, we propose the following methodology based on cross-validation[8]. For each estimation window of 100 months, we select the first 80 months to compute the precision matrices and leave the last 20 observations to minimize the corresponding portfolio variance over the penalty and root parameters. Because this procedure it time consuming, as in Goto and Xu (2013) we apply this procedure in the first estimation window and then we fix the calibrated parameters along the rest of the out-of-sample period. Finally, we also consider different versions of the r-glasso procedure where the root k is fixed from 1 to 5 with increments of 0.5.

Table 4 shows the out-of-sample variances for the different portfolios.

Table 4. The out-of-sample variances for different portfolios.

Method	$p = 80$	$p = 200$	$p = 300$
glasso	0.00203	0.00143	0.00106
r-glasso $k = 1.5$	0.00157	0.00101	0.00103
r-glasso $k = 2$	0.00142	0.00091	0.00088
r-glasso $k = 2.5$	0.00141	0.00088	0.00090
r-glasso $k = 3$	0.00138	0.00229	0.00110
r-glasso $k = 3.5$	0.00155	0.00116	0.00106
r-glasso $k = 4$	0.00158	0.00168	0.00103
r-glasso $k = 4.5$	0.00161	0.00282	0.00100
r-glasso $k = 5$	0.00165	0.00462	0.00108

The results show that the r-glasso method provides lower out-of-sample portfolio risk than that of the glasso method, especially for values of k around 2. Note also larger the portfolio the better to decrease the power of the root k. Again, a good choice for root parameter k could be 2 or 3.

7 Conclusions

In this paper, we provide a new approach for estimating high-dimensional precision matrices, using the ℓ_1 penalization framework. The proposed method is a simple modification of the popular glasso approach based on performing a k-root transformation of the sample covariance matrix. This transformation allows to spread less the corresponding eigenvalues and maintains the converge rate. Through an extensive analysis using both simulated and real data sets, we

[8]In this application, we do not calibrate the parameters using the BIC criterion because the sparsity pattern of the precision matrix does not have an important role

check the proposed improvement helps to attain better performance with no additional costs. In particular, the propose r-glasso method provides lower statistical losses and higher accuracy for covariance selection (e.g. prediction of Gaussian Graphical Models), than those for the glasso method. The proposed method requieres the calibration of an additional parameter k associated with the root transformation. We propose an efficient calibration procedure based on the BIC criterion, but our results also show that the square root transformation ($k = 2$) may be a reasonable choice in practice.

Appendix

A Analytical Proofs

Proof of Theorem 1: We use the following result from Rothman et al. (2008).

Theorem 2. *Let $Z = \{(i,j) : \omega_{ij} \neq 0\}$ and card$(Z) \leq s$. Under the assumptions $A1, A2$ (see section 3), if $\nu \asymp \sqrt{\frac{\log(p)}{n}}$, the following holds.*

$$||\hat{\Omega}_{glasso} - \Omega||_2 = O_P\left(\sqrt{\frac{(p+s)\log(p)}{n}}\right). \tag{16}$$

Proof. For a detailed proof of the Theorem 2 see Rothman et al. (2008). Although the model, provided by these authors, do not penalize the diagonal elements of the precision matrix (the regularization is done through the term $||\Omega^-||_1$, where $\Omega^- = \Omega - \text{diag}(\Omega)$), the same convergence rate can be obtained for model (1), by brief changes in the proof, provided by these authors.

From the problem (4) we can see that $\hat{\Omega}_{\text{r-glasso}} = \hat{\Gamma}^k$. On the other hand the solution $\hat{\Gamma}$ can be considered as the glasso estimator for the matrix $\Omega^{1/k}$, thus the convergence rate of the $\hat{\Gamma}$ is straightforward. Consider the following conditions for the true model:

B1 : $\lambda_{min}(\Omega^{1/k}) \geq \underline{\beta} > 0$,

B2 : $\lambda_{max}(\Omega^{1/k}) \leq \bar{\beta}$,

for some positive values $\bar{\beta}$ and $\underline{\beta}$.

Referring to the Theorem 2 we can say that under the conditions $B1, B2$,[9] if $\xi_k \asymp \sqrt{\frac{\log(p)}{n}}$

$$||\hat{\Gamma} - \Omega^{1/k}||_2 = O_P\left(\sqrt{\frac{(p+t)\log(p)}{n}}\right), \tag{17}$$

where $t \geq \text{card}(Z)$ and $Z = \{(i,j) : [\Omega^{1/k}]_{ij} \neq 0\}$.

Firstly, let as assume that $k \in \mathbf{N}$. The rate (16) can be written as the following:

$$||\hat{\Omega}_{\text{r-glasso}} - \Omega||_2 = ||\hat{\Gamma}^k - \Omega||_2 = ||(\hat{\Gamma} - \Omega^{\frac{1}{k}})(\hat{\Gamma}^{k-1} + ... + \Omega^{\frac{k-1}{k}})||_2. \tag{18}$$

For any matrices A and B the following two inequalities hold

$$||AB||_2 \leq ||A||_2||B||_{spec}, \tag{19}$$

$$||AB||_2 \geq ||A||_2||B||_{min}, \tag{20}$$

where $||B||_{min} = \lambda_{min}(B)$. Using these inequalities and the equality (18), we will have the following:

$$||\hat{\Omega}_{\text{r-glasso}} - \Omega||_2 \leq ||\hat{\Gamma} - \Omega^{\frac{1}{k}}||_2||\hat{\Gamma}^{k-1} + ... + \Omega^{\frac{k-1}{k}}||_{spec}, \tag{21}$$

$$||\hat{\Omega}_{\text{r-glasso}} - \Omega||_2 \geq ||\hat{\Gamma} - \Omega^{\frac{1}{k}}||_2||\hat{\Gamma}^{k-1} + ... + \Omega^{\frac{k-1}{k}}||_{min}. \tag{22}$$

[9]Note that the conditions $A1, A2$ and $B1, B2$ are equivalent, respectively.

From the assumptions $B1$ and $B2$ it follows that $||\Omega^{1/k}||_{min} = O(1)$ and $||\Omega^{1/k}||_{spec} = O(1)$ correspondingly. Using the rate (17) we can write $||\hat{\Gamma}||_{min} = O_P(1)$ and $||\hat{\Gamma}||_{spec} = O_P(1)$. Therefore, since k is finite, we will have the following:

$$||\hat{\Gamma}^{k-1} + ... + \Omega^{\frac{k-1}{k}}||_{spec} = O_P(1), \tag{23}$$

$$||\hat{\Gamma}^{k-1} + ... + \Omega^{(k-1)/k}||_{min} = O_P(1). \tag{24}$$

Finally, from the inequalities (21), (22) and the rates (17), (23), (24) it follows that

$$||\hat{\Omega}_{\text{r-glasso}} - \Omega||_2 = O_P\left(\sqrt{\frac{(p+t)\log(p)}{n}}\right). \tag{25}$$

Note that neither the assumptions $A1, A2$ nor $B1, B2$ give us any information about s and t. Moreover, the sparsity pattern of the matrix Ω does not imply the same pattern for the matrix $\Omega^{1/k}$, and vice-versa. However we can assume that both s and t have the same order, that is $t \asymp s$. Under this assumption we establish the rate (5).

Remark: In the proof of the theorem, given above, it was assumed the parameter k is an integer number. Now assume that k is a rational number, thus we can express k as a fraction $\frac{r}{m}$, where $r, m \in \mathbf{N}$. We can write $\hat{\Omega}_{\text{r-glasso}} = \hat{\Gamma}^{r/m}$. The rate (17) can be written as

$$||\hat{\Gamma} - \Omega^{m/r}||_2 = O_P\left(\sqrt{\frac{(p+t)\log(p)}{n}}\right), \tag{26}$$

and we have the following:

$$||\hat{\Gamma}^r - \Omega^m||_2 = ||(\hat{\Gamma} - \Omega^{\frac{m}{r}})(\hat{\Gamma}^{r-1} + \hat{\Gamma}^{r-2}\Omega^{\frac{m}{r}} + ... + \Omega^{m\frac{r-1}{r}})||_2. \tag{27}$$

Using the inequalities (19) and (20), we can write the following:

$$||\hat{\Gamma}^r - \Omega^m||_2 \leq ||\hat{\Gamma} - \Omega^{\frac{m}{r}}||_2 ||\hat{\Gamma}^{r-1} + \hat{\Gamma}^{r-2}\Omega^{\frac{m}{r}} + ... + \Omega^{m\frac{r-1}{r}}||_{spec}, \tag{28}$$

$$||\hat{\Gamma}^r - \Omega^m||_2 \geq ||\hat{\Gamma} - \Omega^{\frac{m}{r}}||_2 ||\hat{\Gamma}^{r-1} + \hat{\Gamma}^{r-2}\Omega^{\frac{m}{r}} + ... + \Omega^{m\frac{r-1}{r}}||_{min}. \tag{29}$$

Since r and m are finite, using the rates $||\hat{\Gamma}||_{min} = O_P(1)$ and $||\hat{\Gamma}||_{spec} = O_P(1)$ we can write the following :

$$||\hat{\Gamma}^{r-1} + \hat{\Gamma}^{r-2}\Omega^{\frac{m}{r}} + ... + \Omega^{m\frac{r-1}{r}}||_{spec} = O_P(1), \tag{30}$$

$$||\hat{\Gamma}^{r-1} + \hat{\Gamma}^{r-2}\Omega^{\frac{m}{r}} + ... + \Omega^{m\frac{r-1}{r}}||_{min} = O_P(1). \tag{31}$$

From the inequalities (28), (29) and the rates (26), (30), (31) it follows that

$$||\hat{\Gamma}^r - \Omega^m||_2 = O_P\left(\sqrt{\frac{(p+t)\log(p)}{n}}\right). \tag{32}$$

We can rewrite the left-hand side of the rate (32) as follows:

$$||\hat{\Gamma}^r - \Omega^m||_2 = ||\hat{\Gamma}^{m\frac{r}{m}} - \Omega^m||_2 = ||(\hat{\Gamma}^{\frac{r}{m}} - \Omega)(\hat{\Gamma}^{(m-1)\frac{r}{m}} + \hat{\Gamma}^{(m-2)\frac{r}{m}}\Omega + ... + \Omega^{m-1})||_2. \tag{33}$$

Again using the inequalities (19) and (20), we can write the following:

$$||\hat{\Gamma}^r - \Omega^m||_2 \leq ||\hat{\Gamma}^{\frac{r}{m}} - \Omega||_2 ||\hat{\Gamma}^{(m-1)\frac{r}{m}} + \hat{\Gamma}^{(m-2)\frac{r}{m}}\Omega + ... + \Omega^{m-1}||_{spec}, \tag{34}$$

$$||\hat{\Gamma}^r - \Omega^m||_2 \geq ||\hat{\Gamma}^{\frac{r}{m}} - \Omega||_2 ||\hat{\Gamma}^{(m-1)\frac{r}{m}} + \hat{\Gamma}^{(m-2)\frac{r}{m}}\Omega + ... + \Omega^{m-1}||_{min}. \tag{35}$$

As before, we have

$$||\hat{\Gamma}^{(m-1)\frac{r}{m}} + \hat{\Gamma}^{(m-2)\frac{r}{m}}\Omega + ... + \Omega^{m-1}||_{spec} = O_P(1), \tag{36}$$

$$||\hat{\Gamma}^{(m-1)\frac{r}{m}} + \hat{\Gamma}^{(m-2)\frac{r}{m}}\Omega + ... + \Omega^{m-1}||_{min} = O_P(1). \tag{37}$$

From the inequalities (34), (35) and the rates (26), (36), (37) it follows that the norms $||\hat{\Gamma}^r - \Omega^m||_2$ and $||\hat{\Gamma}^{\frac{r}{m}} - \Omega||_2$ have the same convergence rate. Thus, the estimation $\hat{\Omega}_{\text{r-glasso}} = \hat{\Gamma}^{\frac{r}{m}}$ has the convergence rate (25).

B Numerical Results

B.1 Precision Matrix Model 1

Figure 2. Average KL loss over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

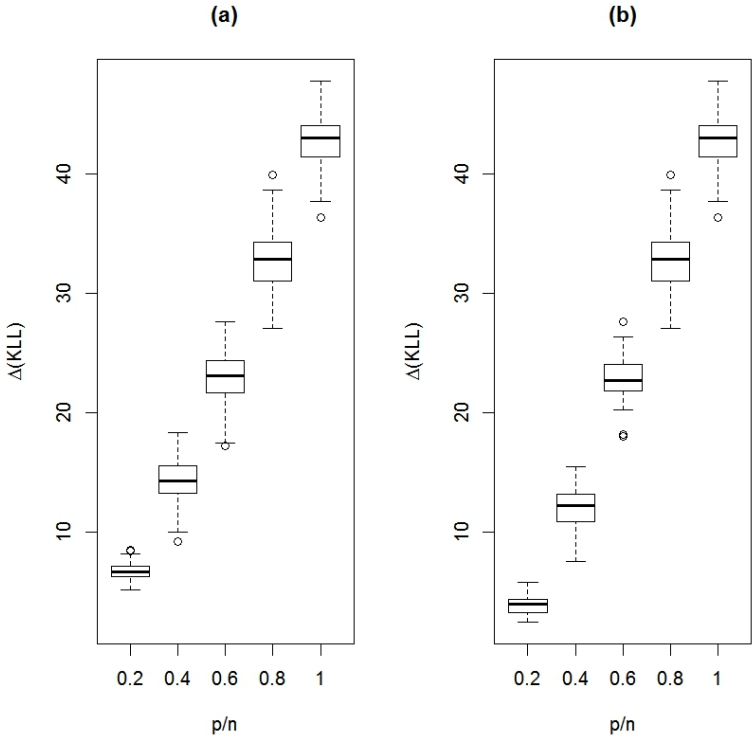


Figure 3. Average MSE over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

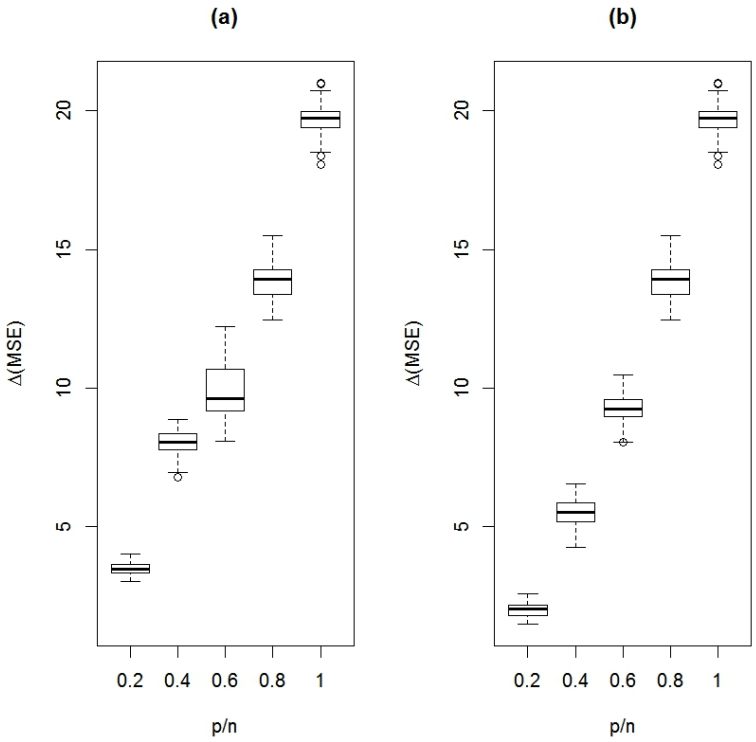


Figure 4. Average Specificity over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

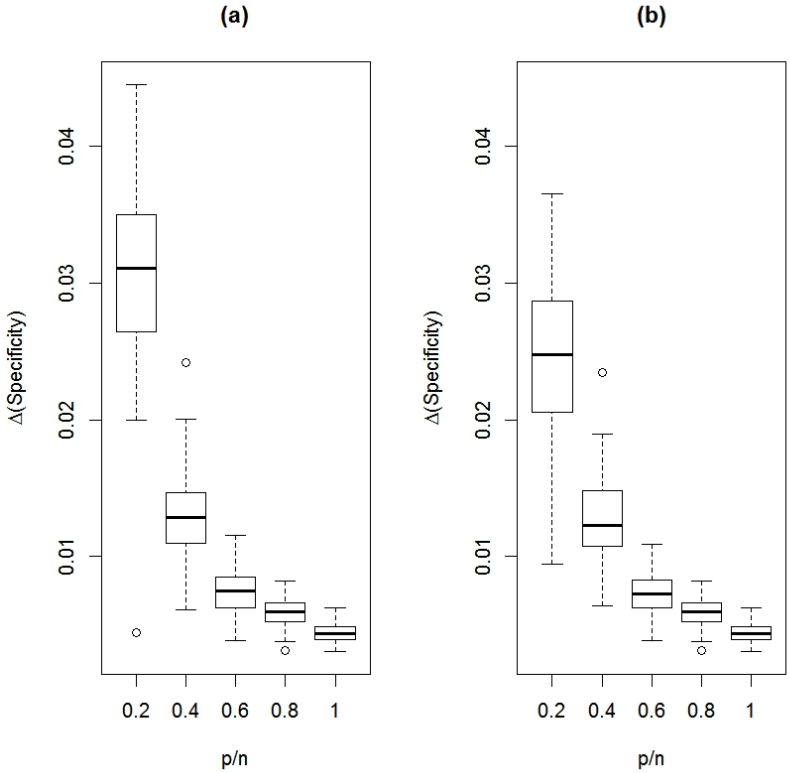


Figure 5. Average Sensitivity over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

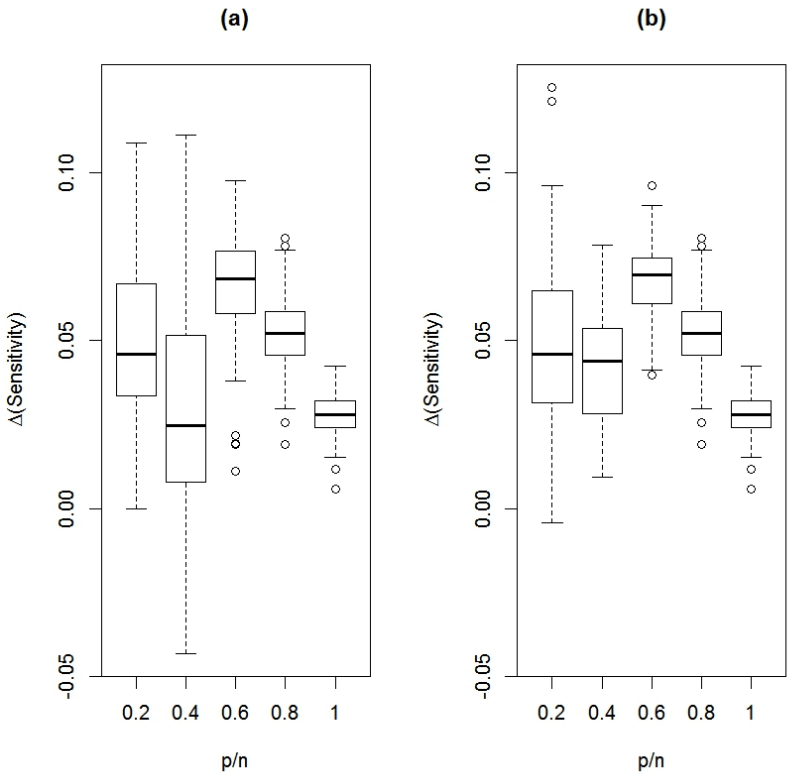
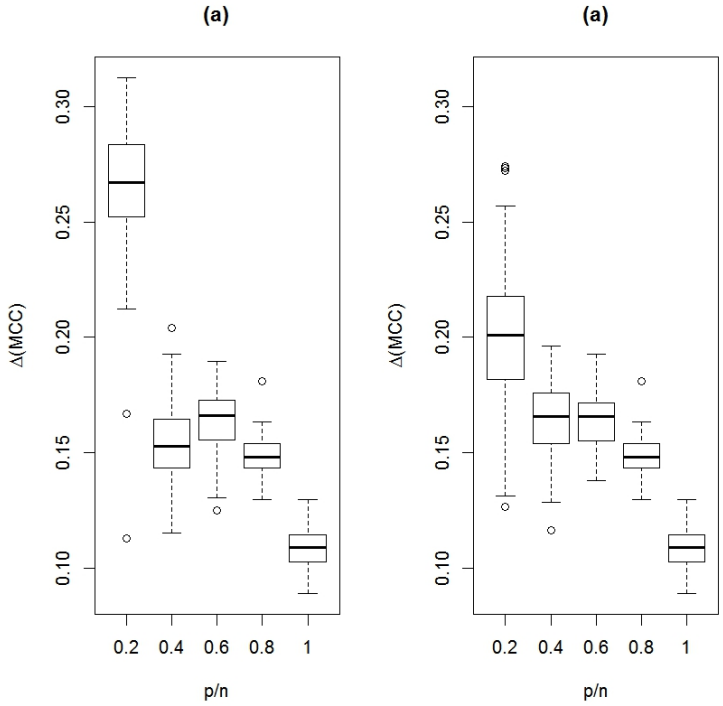


Figure 6. Average MCC over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.



B.2 Precision Matrix Model 2

Figure 7. Average KL loss over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

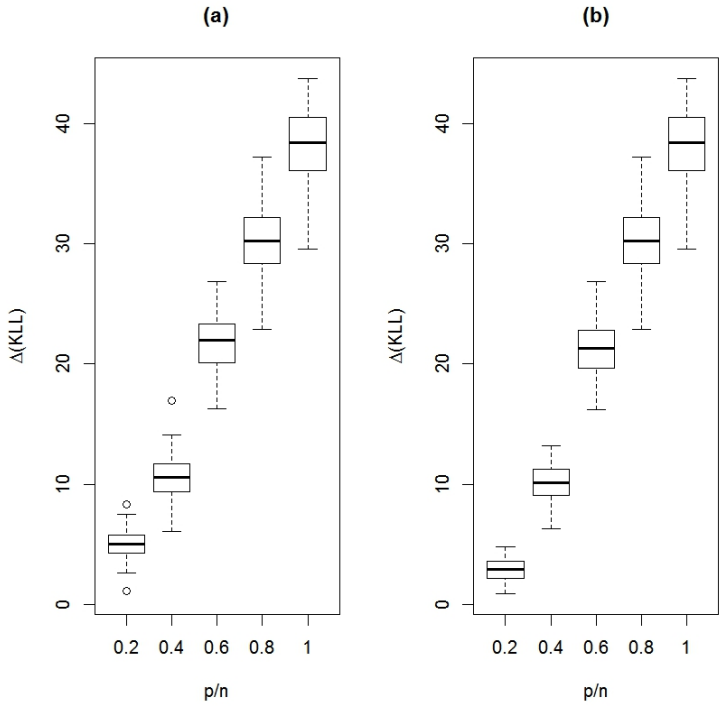


Figure 8. MSE over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

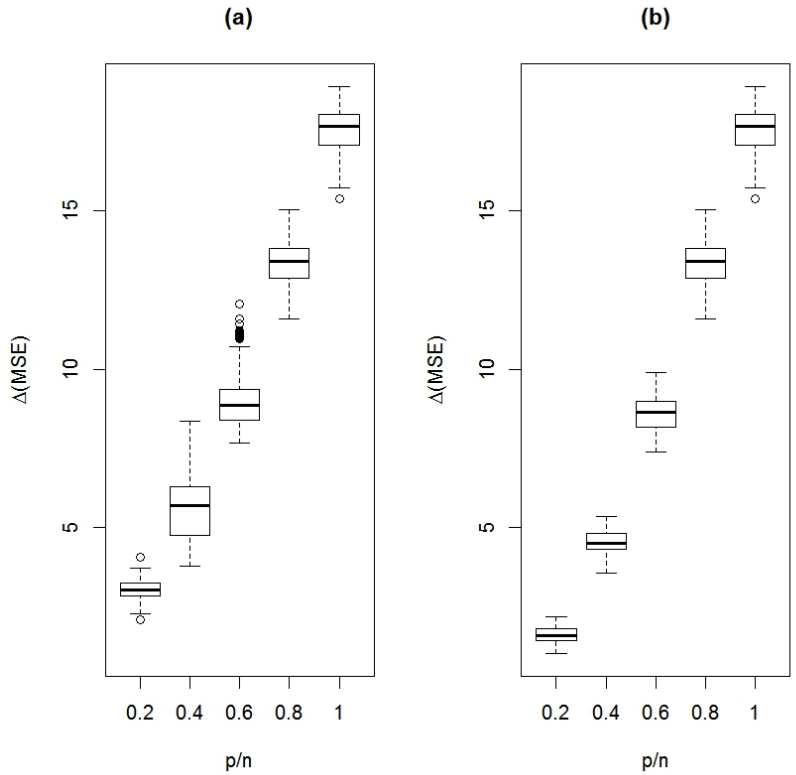


Figure 9. Average Specificity over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

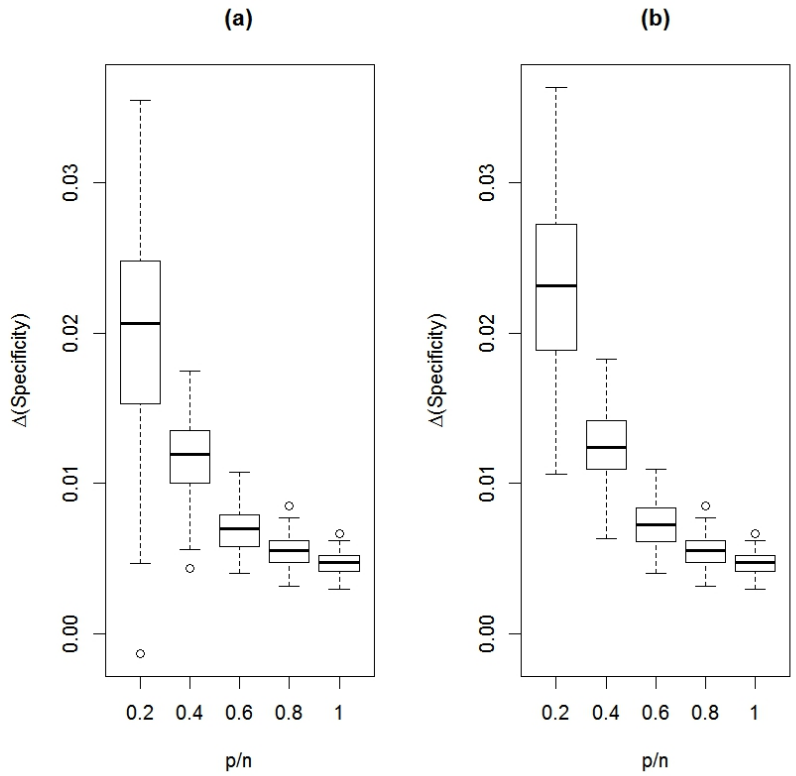


Figure 10. Average Sensitivity over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

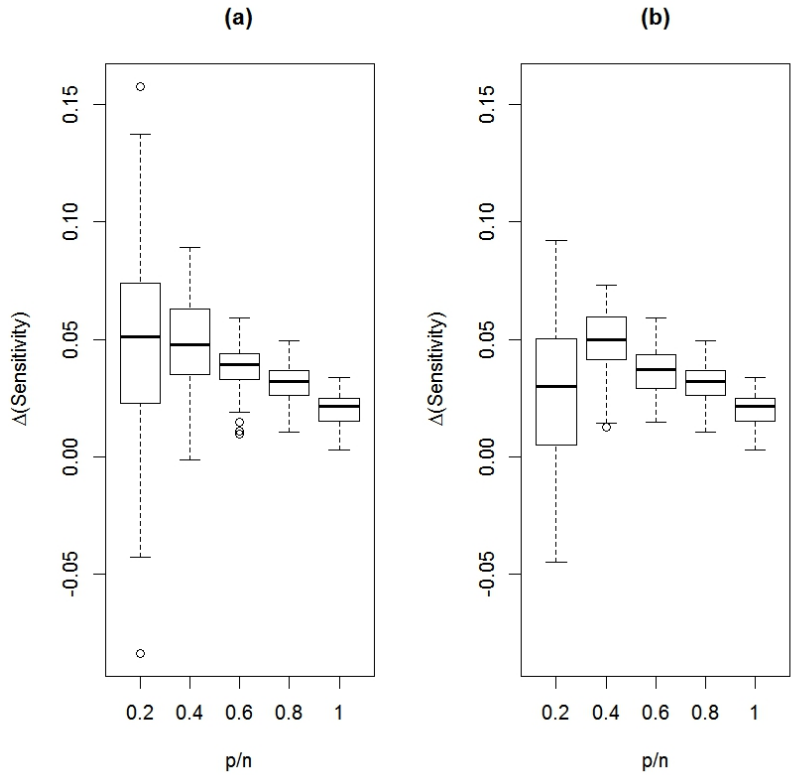
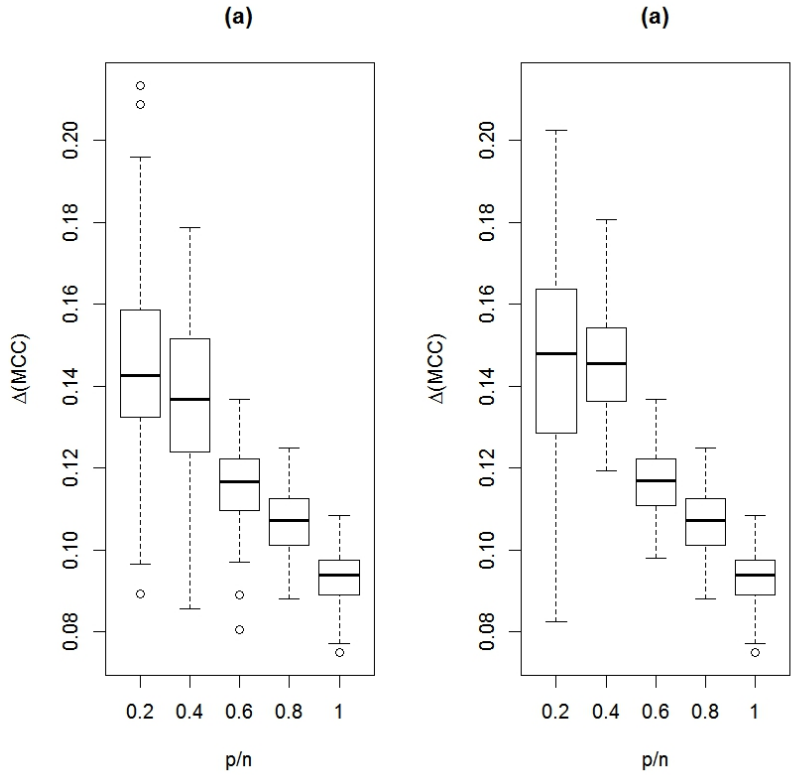


Figure 11. Average MCC over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.



B.3 Precision Matrix Model 3

Figure 12. Average KL loss over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

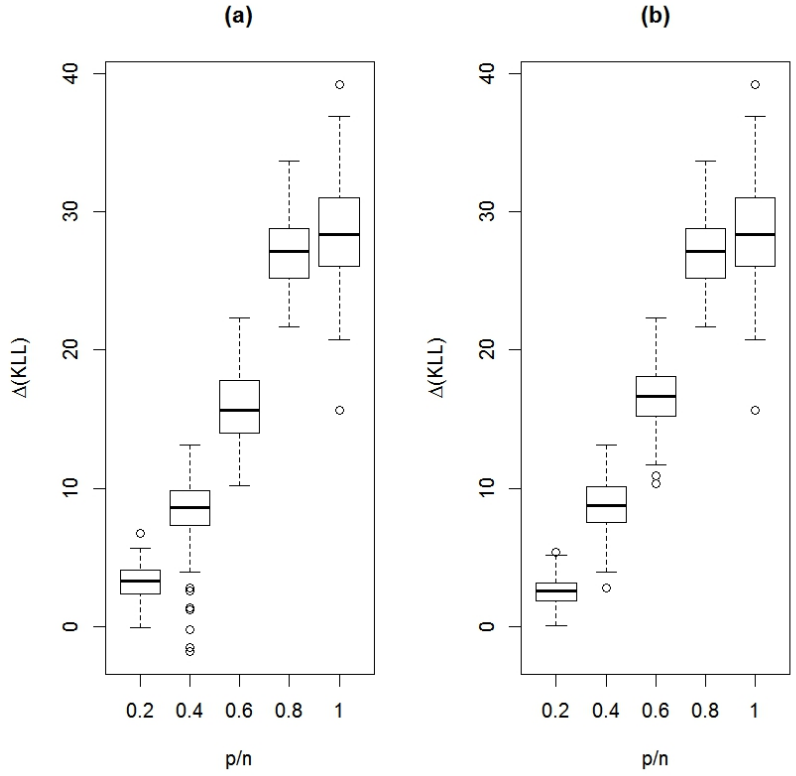


Figure 13. MSE over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

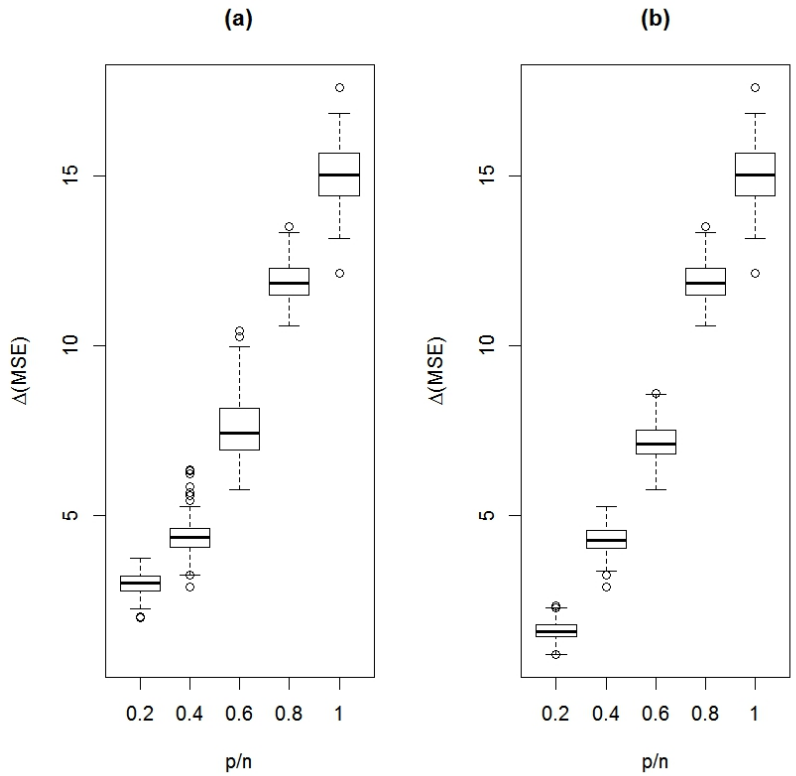


Figure 14. Average Specificity over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

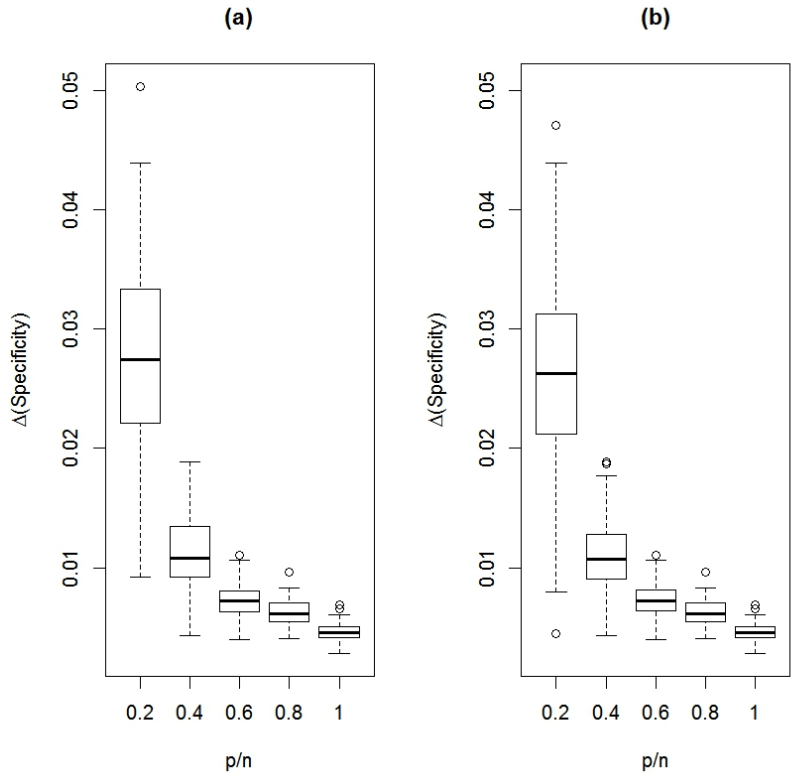


Figure 15. Average Sensitivity over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

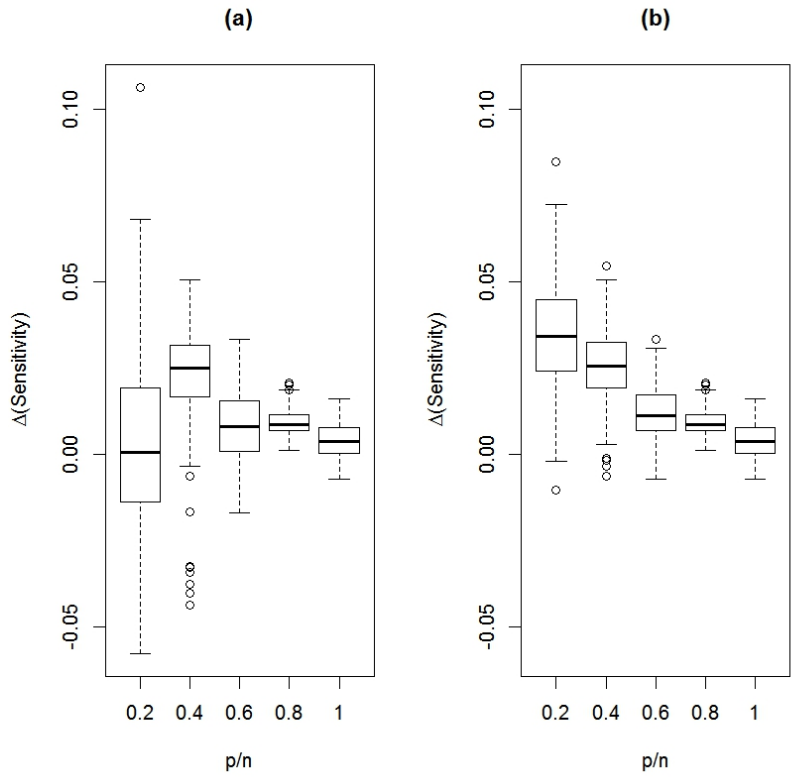
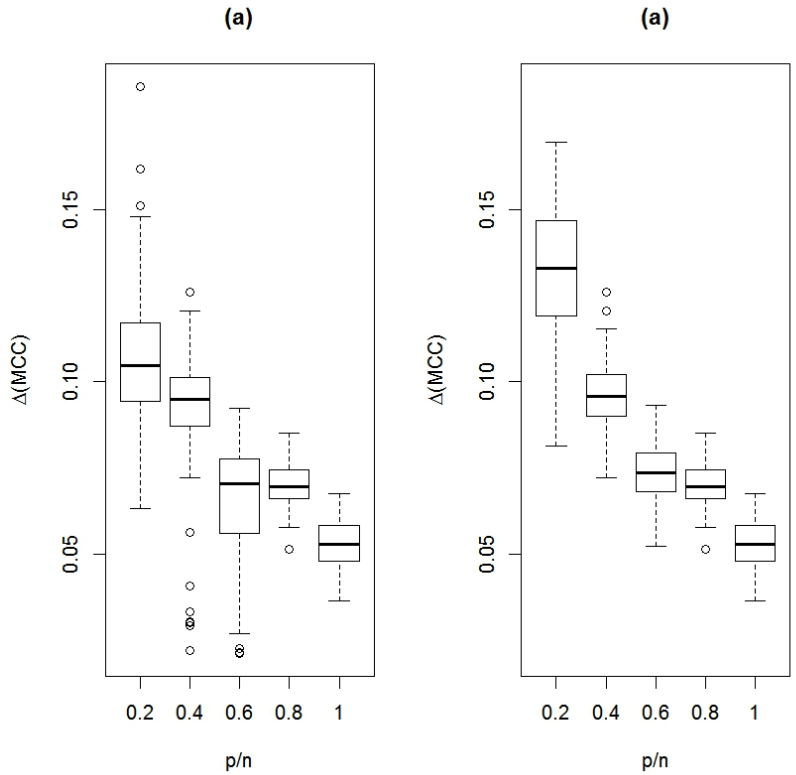


Figure 16. Average MCC over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.



B.4 Precision Matrix Model 4

Figure 17. Average KL loss over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

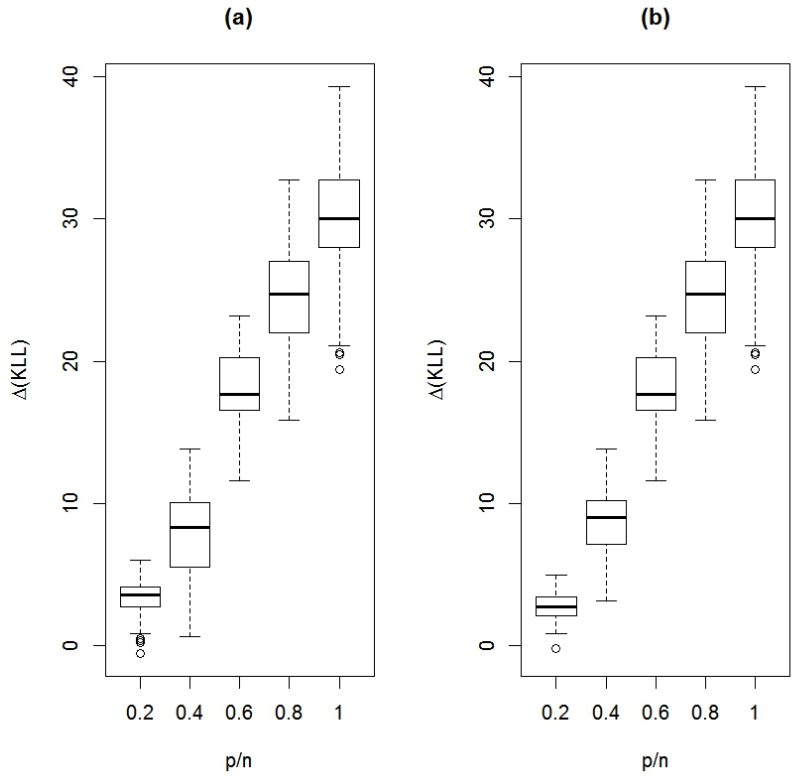


Figure 18. MSE over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

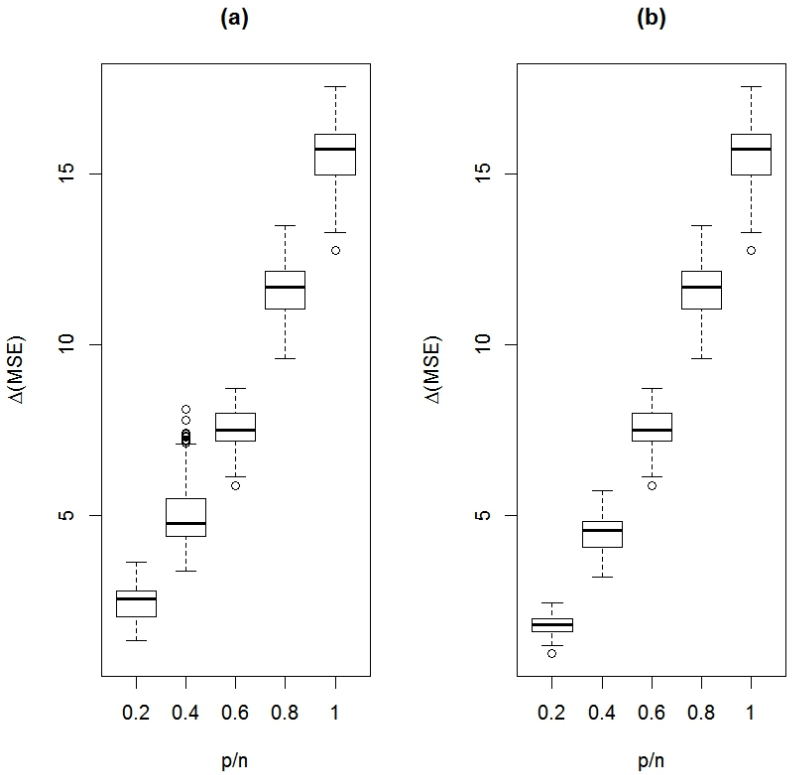


Figure 19. Average Specificity over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

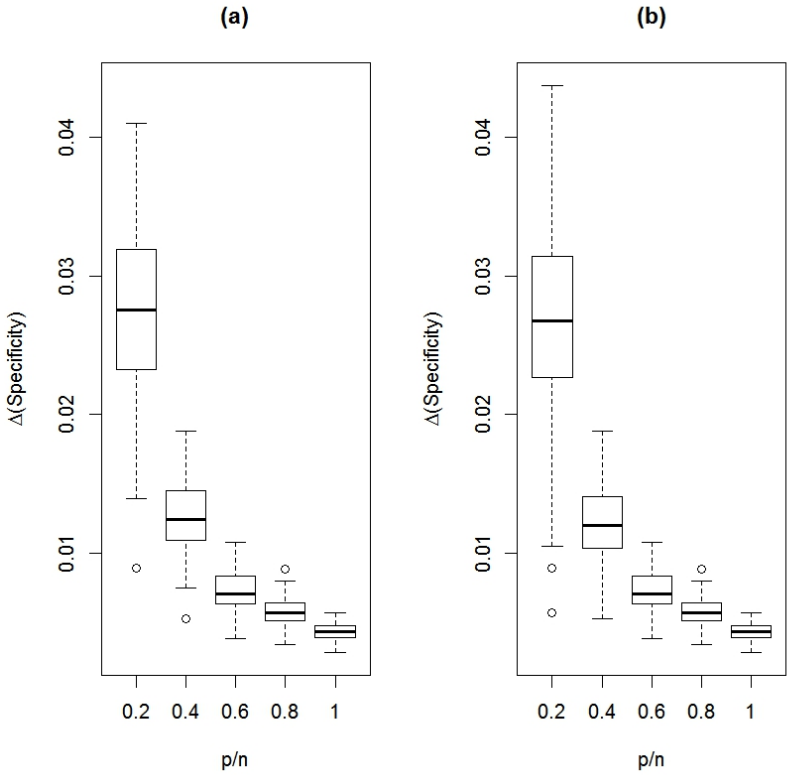


Figure 20. Average Sensitivity over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

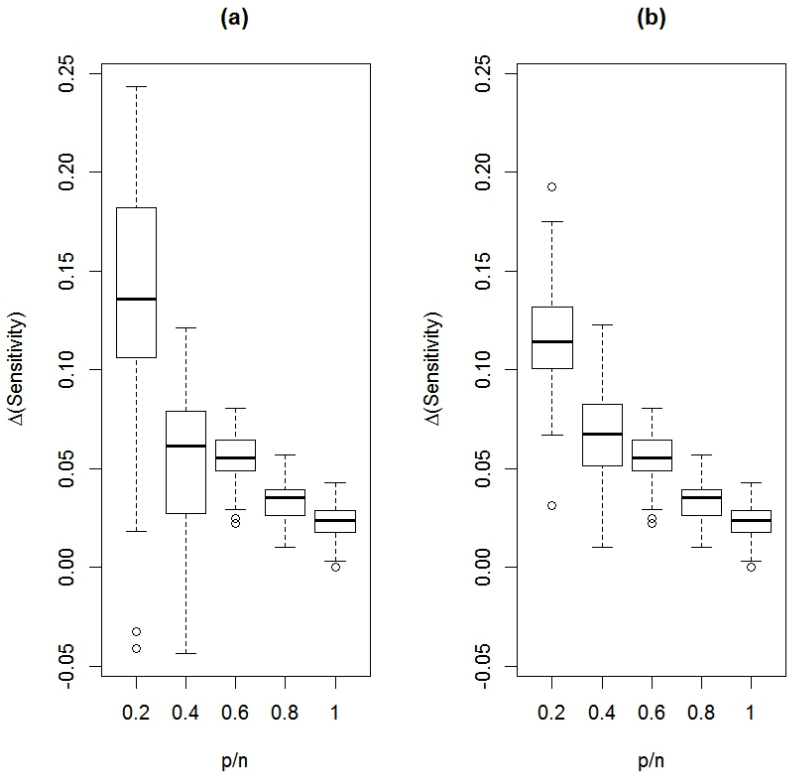
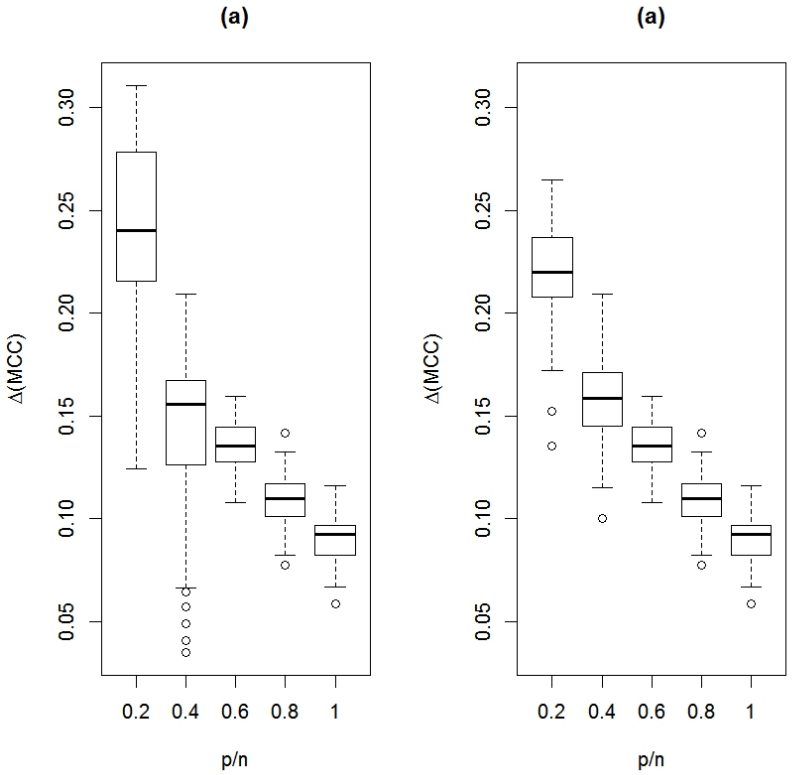


Figure 21. Average MCC over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.



B.5 Precision Matrix Model 5

Figure 22. Average KL loss over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

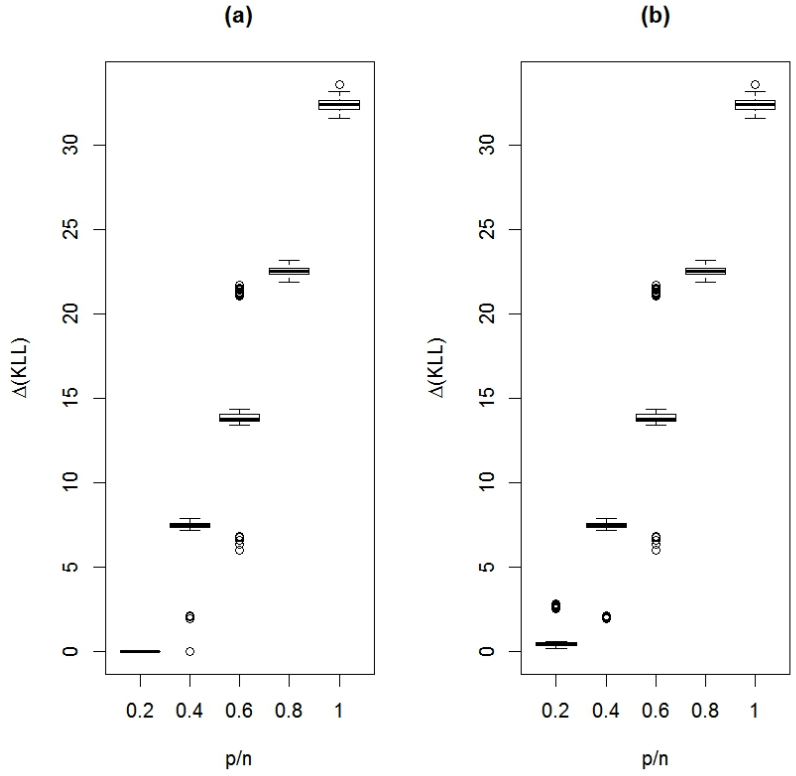


Figure 23. MSE over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

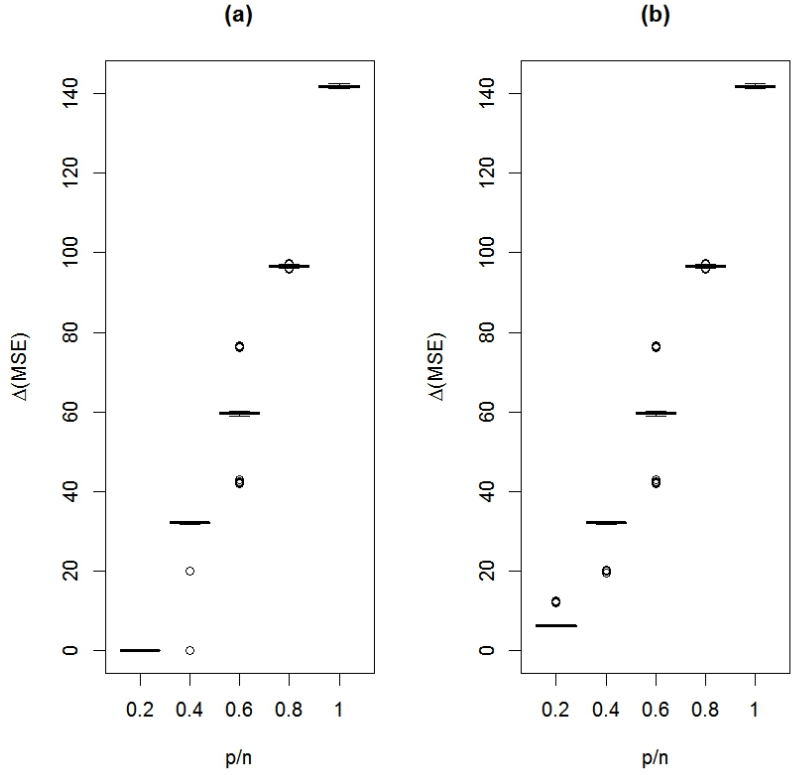


Figure 24. Average Specificity over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

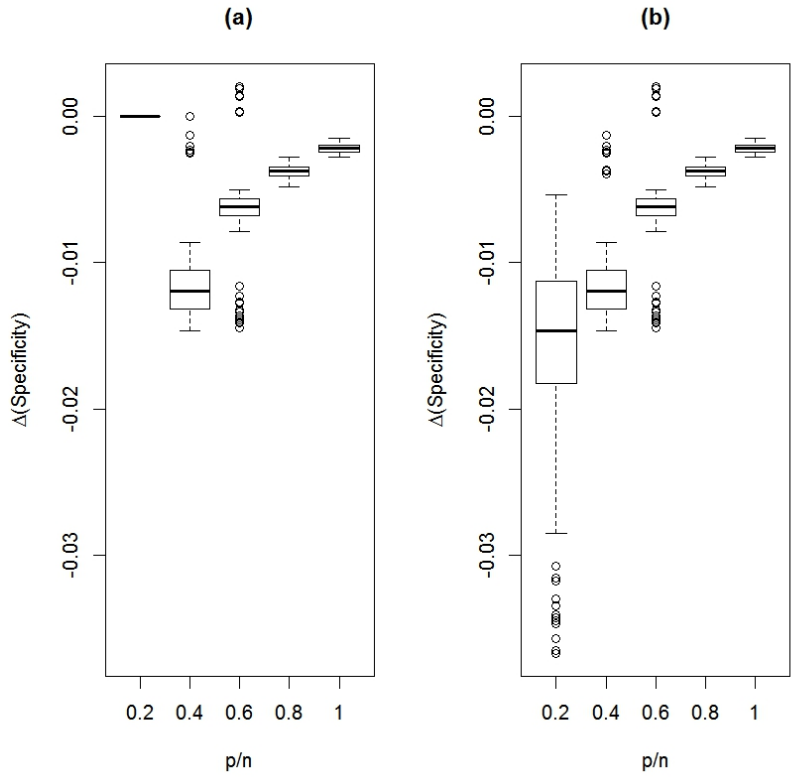


Figure 25. Average Sensitivity over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

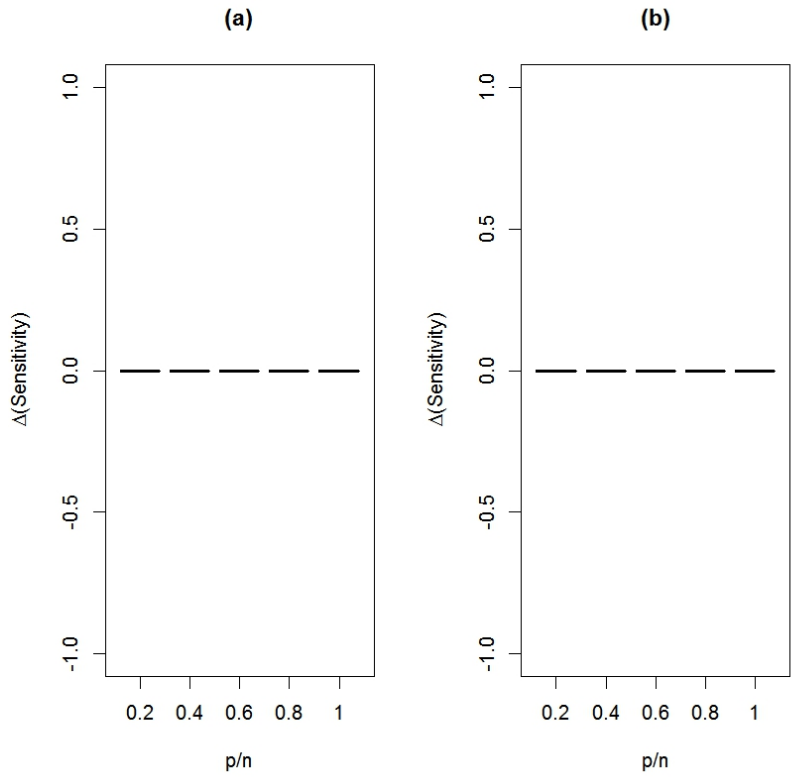
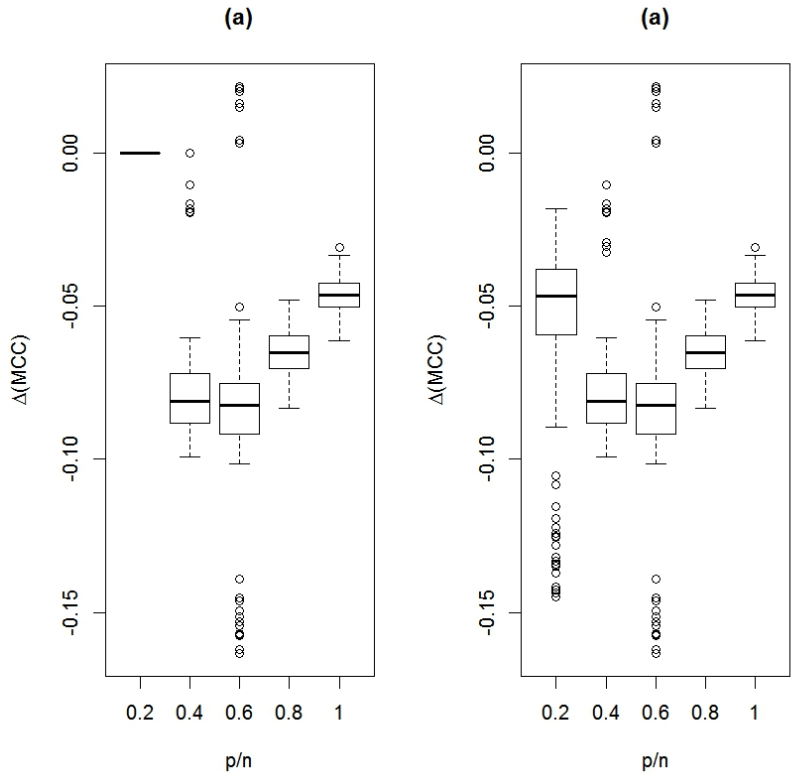


Figure 26. Average MCC over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.



B.6 Precision Matrix Model 6

Figure 27. Average KL loss over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

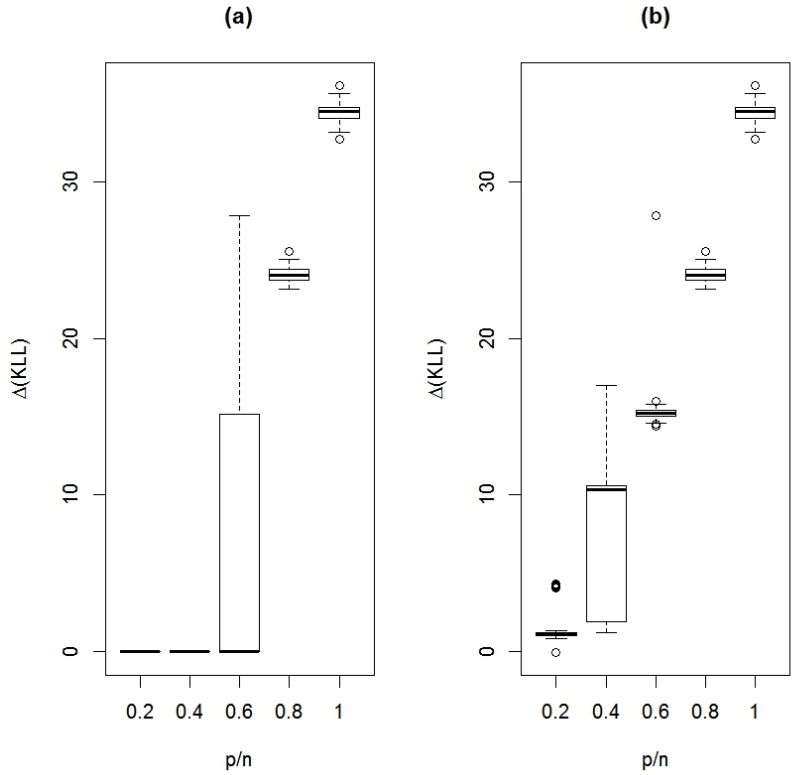


Figure 28. MSE over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

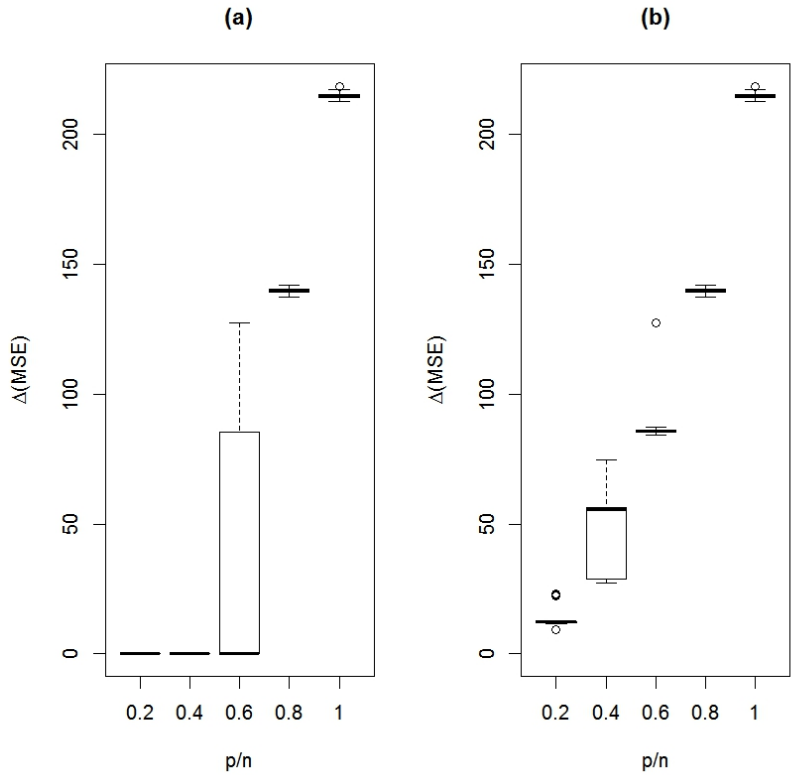


Figure 29. Average Specificity over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

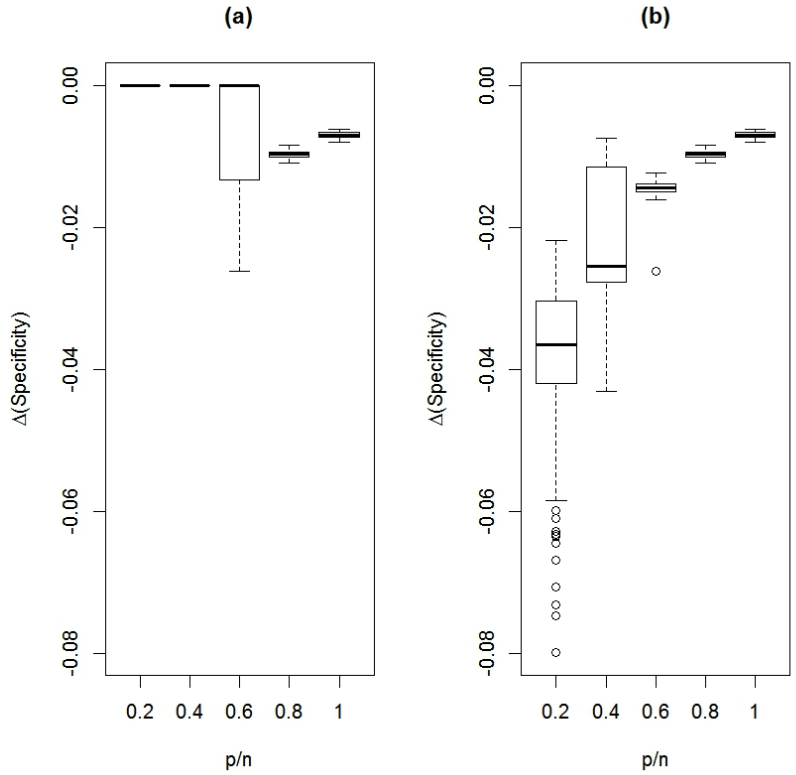


Figure 30. Average Sensitivity over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

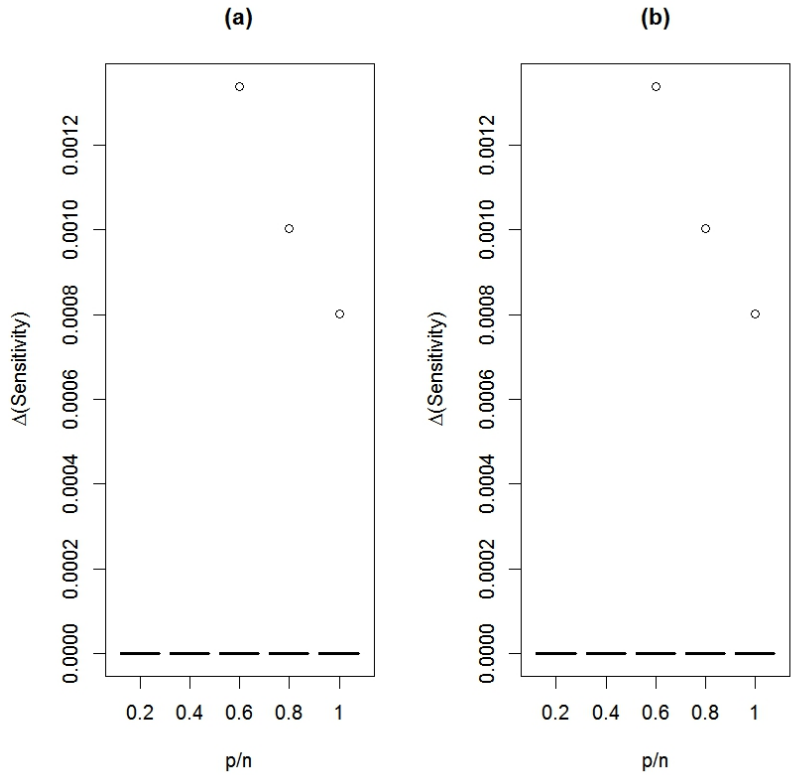
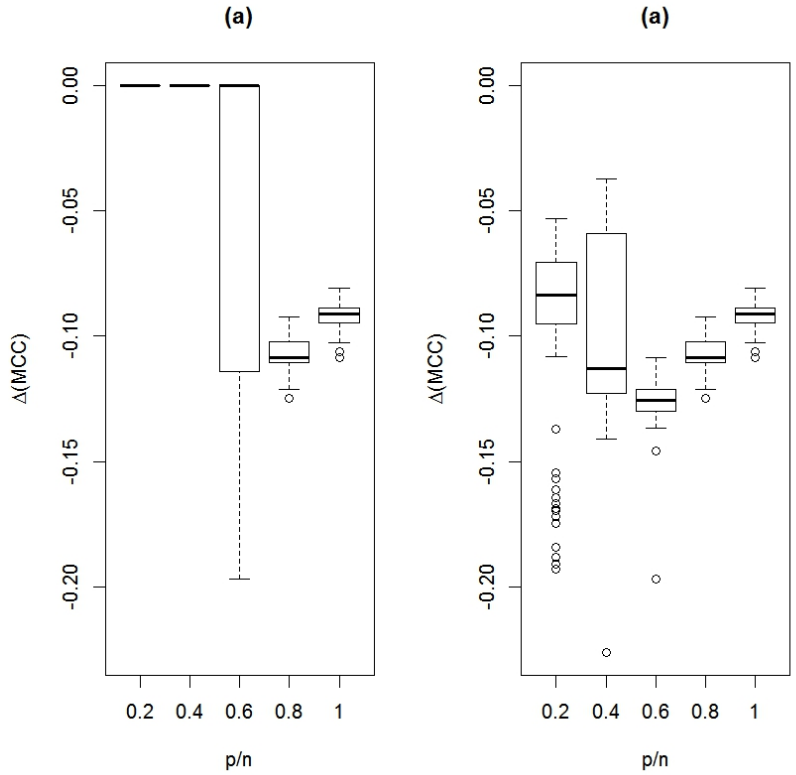


Figure 31. Average MCC over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.



B.7 Precision Matrix Model 7

Figure 32. Average KL loss over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

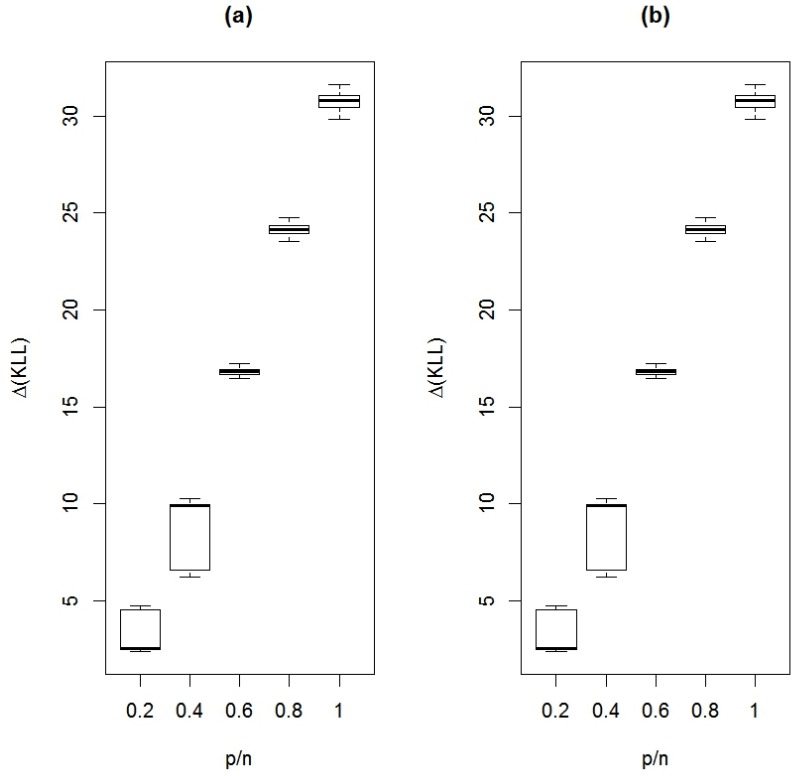


Figure 33. MSE over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

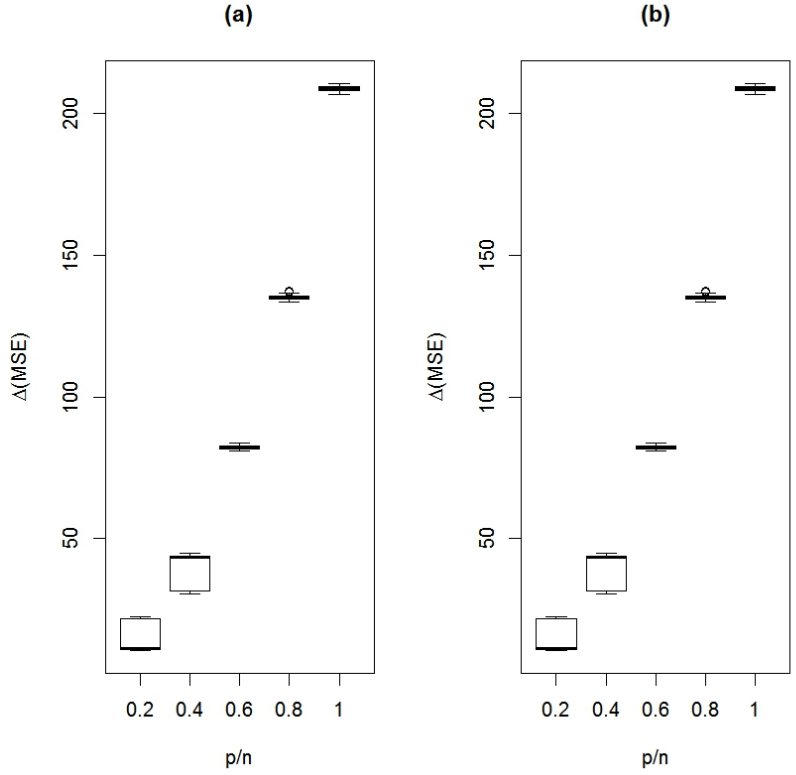
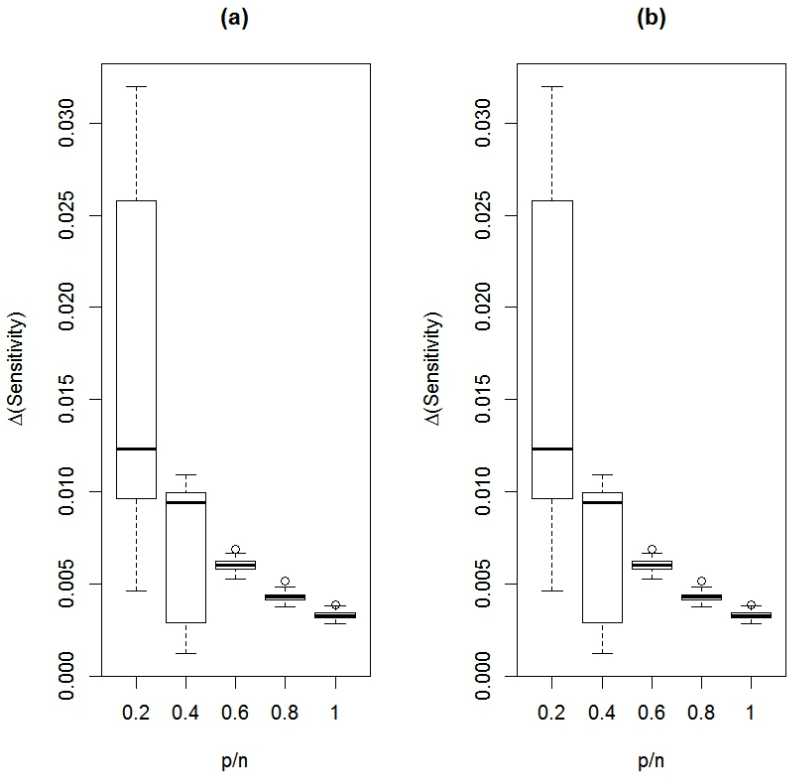


Figure 34. Average Sensitivity over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.



B.8 Precision Matrix Model 8

Figure 35. Average KL loss over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

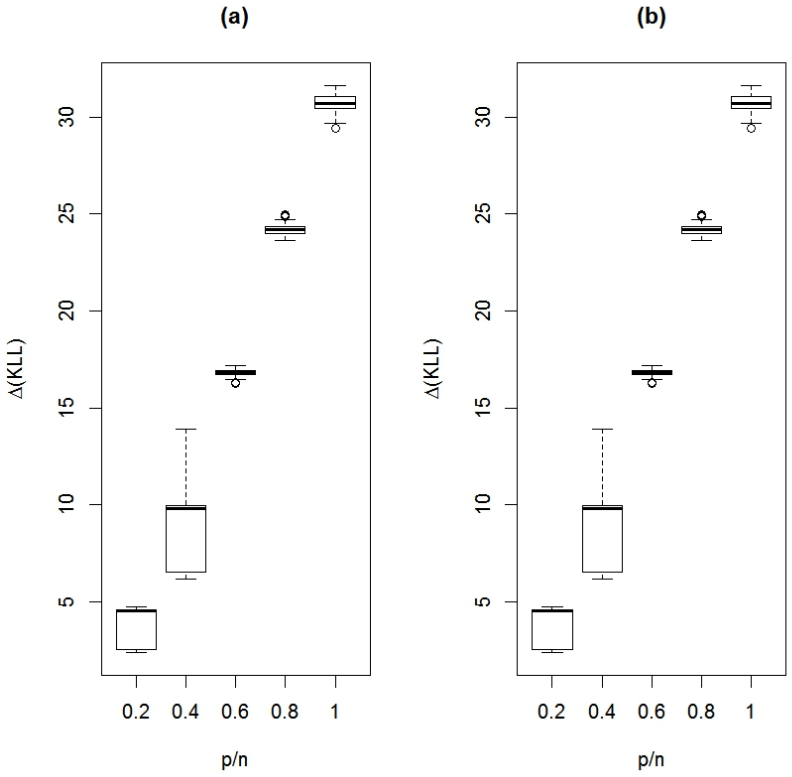


Figure 36. MSE over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

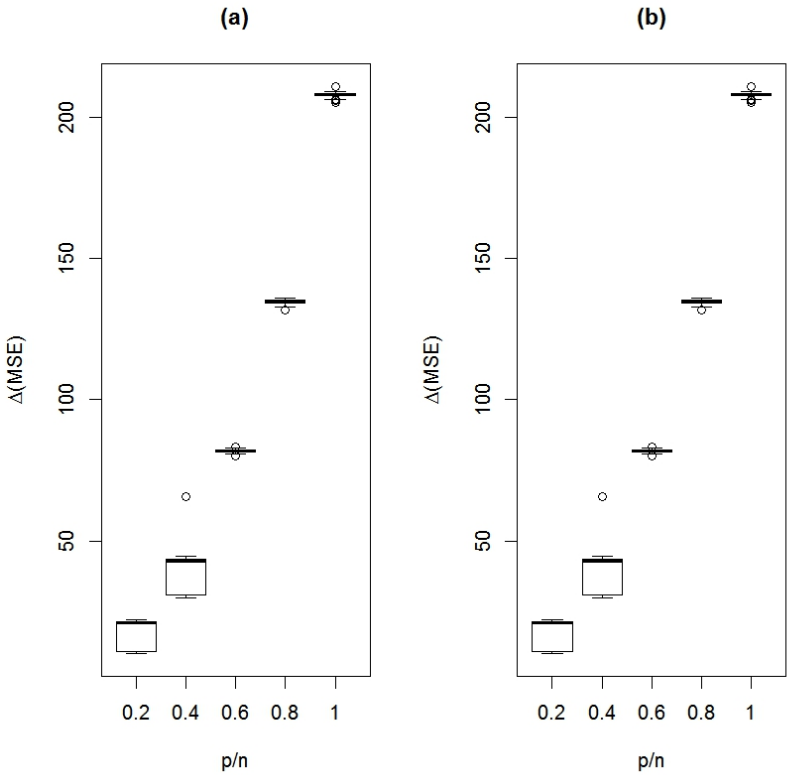


Figure 37. Average Specificity over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

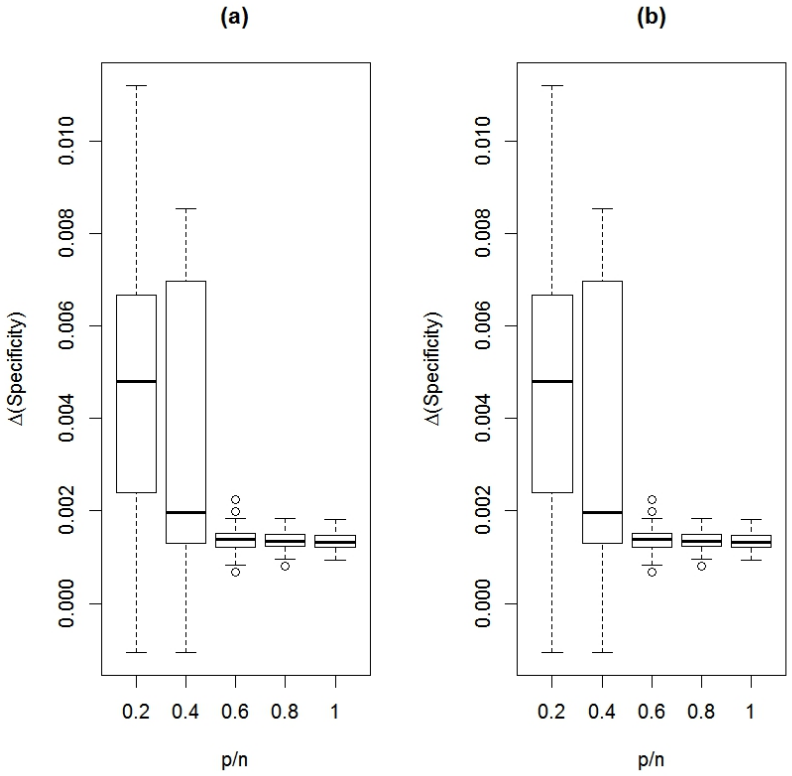


Figure 38. Average Sensitivity over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.

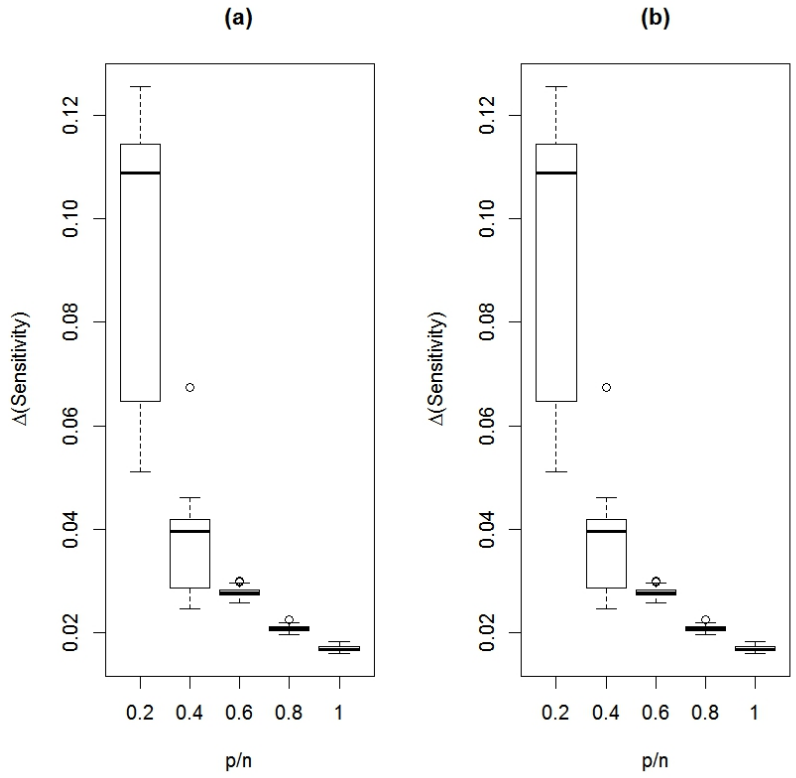
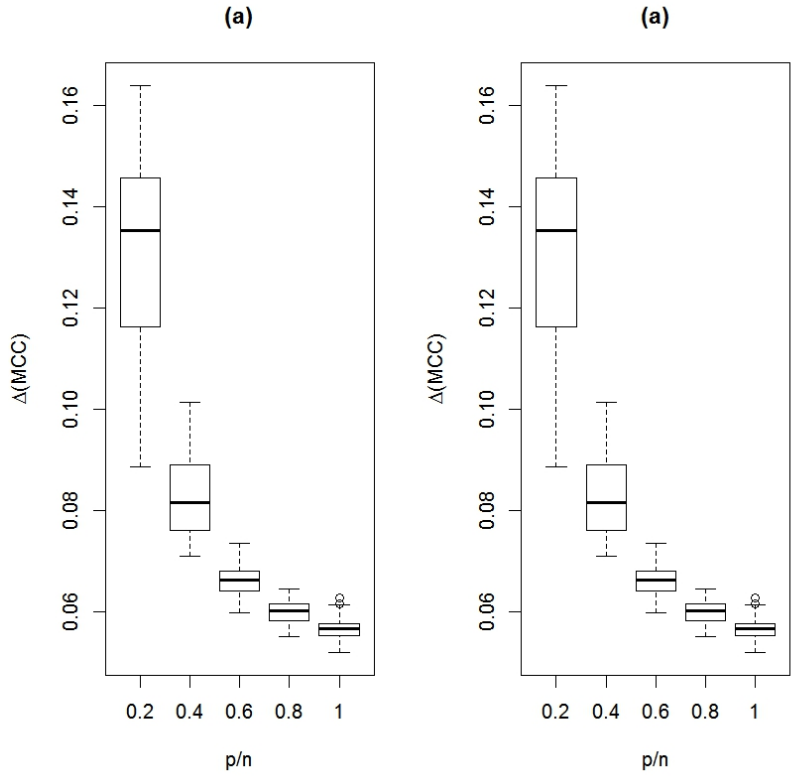


Figure 39. Average MCC over 100 replications for (a) $k = k_{\text{BIC}}$ and (b) $k = 2$.



References

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–172.

Bai, J. and Ng, S. (2011). Principal components estimation and identification of factors. unpublished manuscript.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516.

Banerjee, O., El Ghaoui, L., d'Aspremont, A., and Natsoulis, G. (2006). Convex optimization techniques for fitting sparse gaussian. Pittsburg. Proceeding of the 23^{rd} International Conference on Machine Learning.

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica*, 51:1305–1324.

d'Aspremont, A., Banerjee, O., and Ghaoui, L. (2008). First-order methods for sparse covariance selection. *SIAM J. Matrix Anal Appl.*, 30:56–66.

DeMiguel, V., Garlappi, L., Nogales, J. F., and Uppal, R. (2009). A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55(5):798–812.

Dempster, A. (1972). Covariance selection. *Biometrics*, 28(1):157–175.

Duchi, J., Gould, S., and Koller, D. (2008). Projected subgradient methods for learning sparse gaussians. Proceeding of the Conference on Uncertainty in Artificial Intelligence.

Fan, J., Feng, J., and Wu, Y. (2009). Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics*, 3(2):521–541.

Frahm, G. and Memmel, C. (2010). Dominating estimator for minimum-variance portfolios. *Journal of Econometrics*, 159:289–302.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Goto, S. and Xu, Y. (2013). Improving mean variance optimization through sparse hedging restrictions. *Journal of Financial and Quantitative Analysis, (Forthcoming)*.

Hess, L., Anderson, K., Symmans, W. F., Valero, V., Ibrahim, N., Mejia, J. A., B. D., Theriault, R. L., Buzdar, A. U., Dempsey, P. J., Rouzier, R., Sneige, N., Ross, J. S., Vidaurre, R., Gomez, H. L., Hortobagyi, G. N., and Pusztai, L. (2006). Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of Clinical Oncology*, 24:4236–4244.

Huang, S., Li, J., Sun, L., Ye, J., Fleisher, A., Wu, T., Chen, K., and Reiman, E. (2010). Learning brain connectivity of alzheimer's disease by sparse inverse covariance estimation. *NeuroImage*, 50:935–949.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal component analysis. *The Annals of Statistics*, 29(3):295–327.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, R., A., Peterson, C., and Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7:673–679.

Kourtis, A., Dotsis, G., and Markellos, N. (2012). Parameter uncertainty in portfolio selection: Shrinking the inverse covariance matrix. *Journal of Banking & Finance*, 36:2522–2531.

Kuerer, H. M., Newman, L. A., Smith, T. L., Ames, F. C., Hunt, K. K., Dhingra, K., Theriault, R. L., Singh, G., Binkley, S. M., Sneige, N., Buchholz, T. A., I., R. M., McNeese, M. D., Buzdar, A. U., N., H. G., and Singletary, S. E. (1999). Clinical course of breast cancer patients with complete pathologic primary tumor and axillary lymph node response to doxorubicin-based neoadjuvant chemotherapy. *Journal of Clinical Oncology*, 17(2):460–469.

Lauritzen, S. (1996). *Graphical Models.* Clarendon Press. Oxford.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411.

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060.

Li, L. and Toh, K. (2010). An inexact interior point method for ℓ_1-regurized sparse covariance selection. *Mathematical Programming Computation*, 2:291–315.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis.* Academic Press, New York.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta*, 405:442–451.

McLachlan, S. (2004). *Discriminant Analysis and Statistical Pattern Recognition.* Willey Interscience.

Meinhausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(2):1436–1462.

Rothman, A., Bickel, P., and Levina, E. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186.

Rothman, A., Bickel, P., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

Schafer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1):Article 32.

Scheinberg, K., Ma, S., and Goldfarb, D. (2010). Sparse inverse covariance selection via alternating linearization methods. Proceeding of the Neural Information Processing Systems (NIPS).

Stifanelli, P. F., Creanza, T. M., Anglani, R., Liuzzi, V. C., Mukherjee, S., Schena, F. P., and Ancona, N. (2013). A comparative study of covariance selection models for the inference of gene regulatory networks. *Journal of Biomedical Informatics*, 46:894–904.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288.

Yin, J. and Li, J. (2013). Adjusting for high-dimensional covariates in sparse precision matrix estimation by ℓ_1-penalization. *Journal of Multivariate Analysis*, 116:365–381.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learing Research*, 11:2261–2286.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35.